# Principles and Practice of Explainable Machine Learning

frontiers in Big Data

REVIEW
published: 01 July 2021
doi: 10.3389/fdata.2021.688969

# Principles and Practice of Explainable Machine Learning

*Vaishak Belle*[1,2] *and Ioannis Papantonis*[1]*

[1]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom, [2]Alan Turing Institute, London, United Kingdom

Artificial intelligence (AI) provides many opportunities to improve private and public life. Discovering patterns and structures in large troves of data in an automated manner is a core component of data science, and currently drives applications in diverse areas such as computational biology, law and finance. However, such a highly positive impact is coupled with a significant challenge: how do we understand the decisions suggested by these systems in order that we can trust them? In this report, we focus specifically on data-driven methods—machine learning (ML) and pattern recognition models in particular—so as to survey and distill the results and observations from the literature. The purpose of this report can be especially appreciated by noting that ML models are increasingly deployed in a wide range of businesses. However, with the increasing prevalence and complexity of methods, business stakeholders in the very least have a growing number of concerns about the drawbacks of models, data-specific biases, and so on. Analogously, data science practitioners are often not aware about approaches emerging from the academic literature or may struggle to appreciate the differences between different methods, so end up using industry standards such as SHAP. Here, we have undertaken a survey to help industry practitioners (but also data scientists more broadly) understand the field of explainable machine learning better and apply the right tools. Our latter sections build a narrative around a putative data scientist, and discuss how she might go about explaining her models by asking the right questions. From an organization viewpoint, after motivating the area broadly, we discuss the main developments, including the principles that allow us to study transparent models vs. opaque models, as well as model-specific or model-agnostic post-hoc explainability approaches. We also briefly reflect on deep learning models, and conclude with a discussion about future research directions.

**Keywords: survey, explainable AI, black-box models, transparent models, machine learning**

OPEN ACCESS

***Edited by:***
*Jason Millar,*
*University of Ottawa, Canada*

***Reviewed by:***
*Paolo Giudici,*
*University of Pavia, Italy*
*Doug Talbert,*
*Tennessee Technological University, United States*

***Correspondence:**
*Ioannis Papantonis*
*i.papantonis@sms.ed.ac.uk*

***Specialty section:***
*This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data*

***Received:*** *07 April 2021*
***Accepted:*** *26 May 2021*
***Published:*** *01 July 2021*

***Citation:***
*Belle V and Papantonis I (2021) Principles and Practice of Explainable Machine Learning. Front. Big Data 4:688969. doi: 10.3389/fdata.2021.688969*

## 1 INTRODUCTION

Artificial intelligence (AI) provides many opportunities to improve private and public life. Discovering patterns and structures in large troves of data in an automated manner is a core component of data science, and currently drives applications in diverse areas such as computational biology, law and finance. However, such a highly positive impact is coupled with significant challenges: how do we understand the decisions suggested by these systems in order that we can trust them? Indeed, when one focuses on data-driven methods—machine learning and pattern recognition models in particular—the inner workings of the model can be hard to understand. In the very least, explainability can facilitate the understanding of various

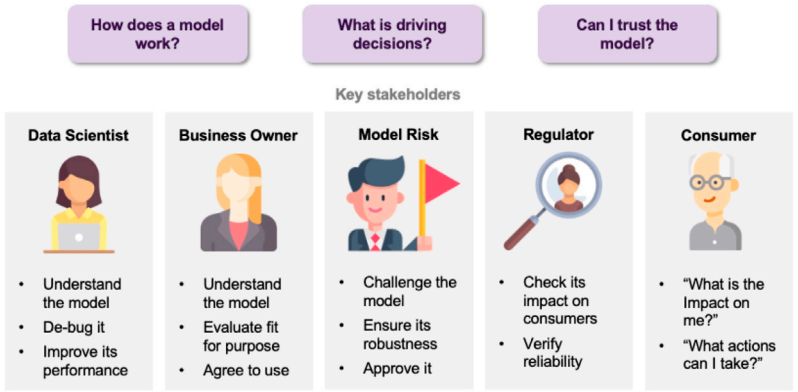


**FIGURE 1 |** Concerns faced by various stakeholders.

aspects of a model, leading to insights that can be utilized by various stakeholders, such as (cf. **Figure 1**):

- **Data scientists** can be benefited when debugging a model or when looking for ways to improve performance.
- **Business owners** caring about the fit of a model with business strategy and purpose.
- **Model Risk analysts** challenging the model, in order to check for robustness and approving for deployment.
- **Regulators** inspecting the reliability of a model, as well as the impact of its decisions on the customers.
- **Consumers** requiring transparency about how decisions are taken, and how they could potentially affect them.

Looking at explainability from another point of view, the developed approaches can help contribute to the following critical concerns that arise when deploying a product or taking decisions based on automated predictions:

- **Correctness:** Are we confident all and only the variables of interest contributed to our decision? Are we confident spurious patterns and correlations were eliminated in our outcome?
- **Robustness:** Are we confident that the model is not susceptible to minor perturbations, but if it is, is that justified for the outcome? In the presence of a missing or noisy data, are we confident the model does not misbehave?
- **Bias:** Are we aware of any data-specific biases that unfairly penalize groups of individuals, and if yes, can we detect and correct them?
- **Improvement:** In what concrete way can the prediction model be improved? What effect would additional training data or an enhanced feature space have?
- **Transferability:** In what concrete way can the prediction model for one application domain be applied to another application domain? What properties of the data and model would have to be adapted for this transferability?
- **Human comprehensibility:** Are we able to explain the model's algorithmic machinery to an expert? Perhaps even a lay person? Is that a factor for deploying the model more widely?

The **purpose** of this report can be especially appreciated by noting that ML models are increasingly deployed in a wide range of businesses. However, with the increasing prevalence and complexity of methods, business stakeholders in the very least have a growing number of concerns about the drawbacks of models, data-specific biases, and so on. Analogously, data science practitioners are often not aware about approaches emerging from the academic literature, or may struggle to appreciate the differences between different methods, so end up using industry standards such as SHAP (Lundberg and Lee, 2017). In this report, we have undertaken a survey to help industry practitioners (but also data scientists more broadly) understand the field of explainable machine learning better and apply the right tools. Our latter sections particularly target how to distill and streamline questions and approaches to explainable machine learning.

## 2 DEVELOPMENT AND CONTRIBUTIONS

Such concerns have motivated intense activity within the community, leading to a number of involved but closely related observations. Drawing on numerous insightful surveys and perspectives [including (Lipton, 2016; Doshi-Velez and Kim, 2017; Arrieta et al., 2019; Weld and Bansal, 2019; Molnar, 2020)] and a large number of available approaches, the goal of this survey is to help shed some light into the various kind of insights that can be gained, when using them. We distill concepts and strategies with the overall aim of helping industry practitioners (but also data scientists more broadly) disentangle the different notions of explanations, as well as their intended scope of application, leading to a better understanding of the field. To this end, we first provide general perspectives on explainable machine learning that covers: notions of transparency, criteria for evaluating explainability, as well as the type of explanations one can expect in general. We then turn to some frameworks for summarizing developments on explainable machine learning.

A taxonomic framework provides an overview of explainable ML, and the other two frameworks study certain aspects of the taxonomy. A detailed discussion on transparent vs. opaque models, model specific vs. model agnostic approaches, as well as post-hoc[1] explainability approaches follows, all of which are referred to in the taxonomic framework. Limitations and strengths of these models and approaches are discussed subsequently. We then turn to brief observations on explainability with respect to deep learning models. Finally, we distill these results further by building a narrative around a putative data scientist, and discuss how she might go about explaining her models. We conclude with some directions for future research, including the need for causality-related properties in machine learning models.

## 3 SCOPE

In the interest of space, we will focus on data-driven methods—machine learning and pattern recognition models in particular—with the primarily goal of classification or prediction by relying on statistical association. Consequently, these engender a certain class of statistical techniques for simplifying or otherwise interpreting the model at hand.

Despite this scoping, the literature is vast.[2] Indeed, we note that underlying concerns about human comprehensibility and generating explanations for decisions is a general issue in cognitive science, social science and human psychology (Miller, 2019). There are also various "meta"-views on explainability, such as maintaining an explicit model of the user (Chakraborti et al., 2019; Kulkarni et al., 2019). Likewise, causality is expected to play a major role in explanations (Miller, 2019), but many models arising in the causality literature require careful experiment design and/or knowledge from an expert (Pearl, 2018). They are, however, an interesting and worthwhile direction for future research, and left for concluding thoughts. Our work here primarily focuses on "mainstream" ML models, and the corresponding statistical explanations (however limiting they may be in a larger context) that one can extract from these models. On that note, we are not concerned with "generating" explanations, which might involve, say, a natural language understanding component, but rather extracting an interpretation of the model's behavior and decision boundary. This undoubtedly limits the literature in terms of what we study and analyze, but it also allows us to be more comprehensive in that scope. For simplicity, we will nonetheless abbreviate this scoping of explainable machine learning as XAI in the report, but reiterate that the AI community takes a broader view that goes beyond (statistical) classification tasks (Chakraborti et al., 2019; Kulkarni et al., 2019).

[1]The term post-hoc reflects the fact that explainability approaches inspect a model after the training is completed, thus they do not influence or interfere with the training process, they only audit the resulting model to assess its quality.

[2]A search on Google Scholar for "explainable machine learning" returns about one thousand results; varying search to disjunctively include terms such as "interpretable," "artificial intelligence," and "explanations," returns an even more extensive set of research papers, naturally.

While we do survey and distill approaches to provide a high-level perspective, we expect the reader to have some familiarity with classification and prediction methods. Finally, in terms of terminology, we will mostly use the term "model" to mean the underlying machine learning technique such as random forests or logistic regression or convolutional neural networks, and use the term "approach" and "method" to mean an algorithmic pipeline that is undertaken to explicitly simplify, interpret or otherwise obtain explanations from a model. If we deviate from this terminology, the context will make clear whether the entity is a machine learning or an explainability one.

## 4 PERSPECTIVES ON EXPLAINABILITY

Before delving into actual approaches for explainability, it is worthwhile to reflect on what are the dimensions for human comprehensibility. We will start with notions of transparency, in the sense of humans understanding the inner workings of the model. We then turn to evaluation criteria for models. We finally discuss the types of explanations that one might desire from models. It should be noted that there is considerable overlap between these notions, and in many cases, a rigorous definition or formalization is lacking and generally hard to agree on.

### 4.1 Transparency

Transparency stands for a human-level understanding of the inner workings of the model (Lipton, 2016). We may consider three dimensions:

- **Simulatability** is the first level of transparency and it refers to a model's ability to be simulated by a human. Naturally, only models that are simple and compact fall into this category. Having said that, it is worth noting that simplicity alone is not enough, since, for example, a very large amount of simple rules would prohibit a human to calculate the model's decision simply by thought. On the other hand, simple cases of otherwise complex models, such as a neural network with no hidden layers, could potentially fall into this category.
- **Decomposability** is the second level of transparency and it denotes the ability to break down a model into parts (input, parameters and computations) and then explain these parts. Unfortunately, not all models satisfy this property.
- **Algorithmic Transparency** is the third level and it expresses the ability to understand the procedure the model goes through to generate its output. For example, a model that classifies instances based on some similarity measure (such as K-nearest neighbors) satisfies this property, since the procedure is clear; find the datapoint that is the most similar to the one under consideration and assign to the former the same class as the latter. On the other hand, complex models, such as neural networks, construct an elusive loss function, while the solution to the training objective has to be approximated, too. Generally speaking, the only requirement for a model to fall into this category is for the user to be able to inspect it through a mathematical analysis.

Broadly, of course, we may think of machine models as either being transparent or opaque/black-box, although the above makes clear this distinction is not binary. In practice, despite the nuances, it is convention to see decision trees, linear regression, among others as simpler, transparent models, and random forests, deep learning, among others as opaque models, partly because current applications rarely use a single perceptron neural network.

## 4.2 Evaluation Criteria

Although initially considered for rule extraction methods (Craven and Shavlik, 1999), we might consider the following dimensions to evaluating models in terms of explainability:

- **Comprehensibility:** The extent to which extracted representations are humanly comprehensible, and thus touching on the dimensions of transparency considered earlier.
- **Fidelity:** The extent to which extracted representations accurately capture the opaque models from which they were extracted.
- **Accuracy:** The ability of extracted representations to accurately predict unseen examples.
- **Scalability:** The ability of the method to scale to opaque models with large input spaces and large numbers of weighted connections.
- **Generality:** The extent to which the method requires special training regimes or restrictions on opaque models.

We reiterate that such concepts are hard to quantify rigorously, but can nonetheless serve as guiding intuition for future developments in the area.

## 4.3 Types of Explanations

For opaque models in particular, we might consider the following types of post-hoc explanations (Arrieta et al., 2019):

- **Text explanations** produce explainable representations utilizing symbols, such as natural language text. Other cases include propositional symbols that explain the model's behavior by defining abstract concepts that capture high level processes.
- **Visual explanation** aim at generating visualizations that facilitate the understanding of a model. Although there are some inherit challenges (such as our inability to grasp more than three dimensions), the developed approaches can help in gaining insights about the decision boundary or the way features interact with each other. Due to this, in most cases visualizations are used as complementary techniques, especially when appealing to a non-expert audience.
- **Local explanations** attempt to explain how a model operates in a certain area of interest. This means that the resulting explanations do not necessarily generalize to a global scale, representing the model's overall behavior. Instead, they typically approximate the model around the instance the user wants to explain, in order to extract explanations that describe how the model operates when encountering such instances.
- **Explanations by example** extract representative instances from the training dataset to demonstrate how the model operates. This is similar to how humans approach explanations in many cases, where they provide specific examples to describe a more general process. Of course, for an example to make sense, the training data has to be in a form that is comprehensible by humans, such as images, since arbitrary vectors with hundreds of variables may contain information that is difficult to uncover.
- **Explanations by simplification** refer to the techniques that approximate an opaque model using a simpler one, which is easier to interpret. The main challenge comes from the fact that the simple model has to be flexible enough so it can approximate the complex model accurately. In most cases, this is measured by comparing the accuracy (for classification problems) of these two models.
- **Feature relevance explanations** attempt to explain a model's decision by quantifying the influence of each input variable. This results in a ranking of importance scores, where higher scores mean that the corresponding variable was more important for the model. These scores alone may not always constitute a complete explanation, but serve as a first step in gaining some insights about the model's reasoning.

We now turn to a distillation of the observations and techniques from the literature in the following section. We will not always be able to cover the entire gamut of dimensions considered in this section, but they do serve as a guide for the considerations to follow.

# 5 EXPLORING EXPLAINABLE MACHINE LEARNING

To summarize the rapid development in explainable machine learning (XAI), we turn to five "frameworks" that summarize or otherwise distill the literature. These frameworks can be thought of as a comparative exposition and/or visualization of sorts, which help us understand:

- the limitations of models that may already be deployed (at least regarding explainability),
- what approaches are available for explaining such models, and
- what models may be considered alternatively if the application were to be redesigned with explainability n mind.

As should be expected, there will be overlap between these frameworks.[3] The first two frameworks are inspired by the discussions in (Arrieta et al., 2019), adapted and modified slightly for our purposes. The third and fourth framework are based on an analysis on the current strengths and limitations of popular realizations of XAI techniques. The fifth is a "cheat sheet"

[3]We note that without experimental comparisons and a proper deliberation on the application domain, these frameworks purely provide an intuitive picture of model capabilities. We also note that in what follows, we make the assumption that the data is already segmented and cleaned, but it should be clear that often data pre-processing is a major step before machine learning methods can be applied. Dealing with data that has not been treated can affect both the applicability and the usefulness of explainability methods.

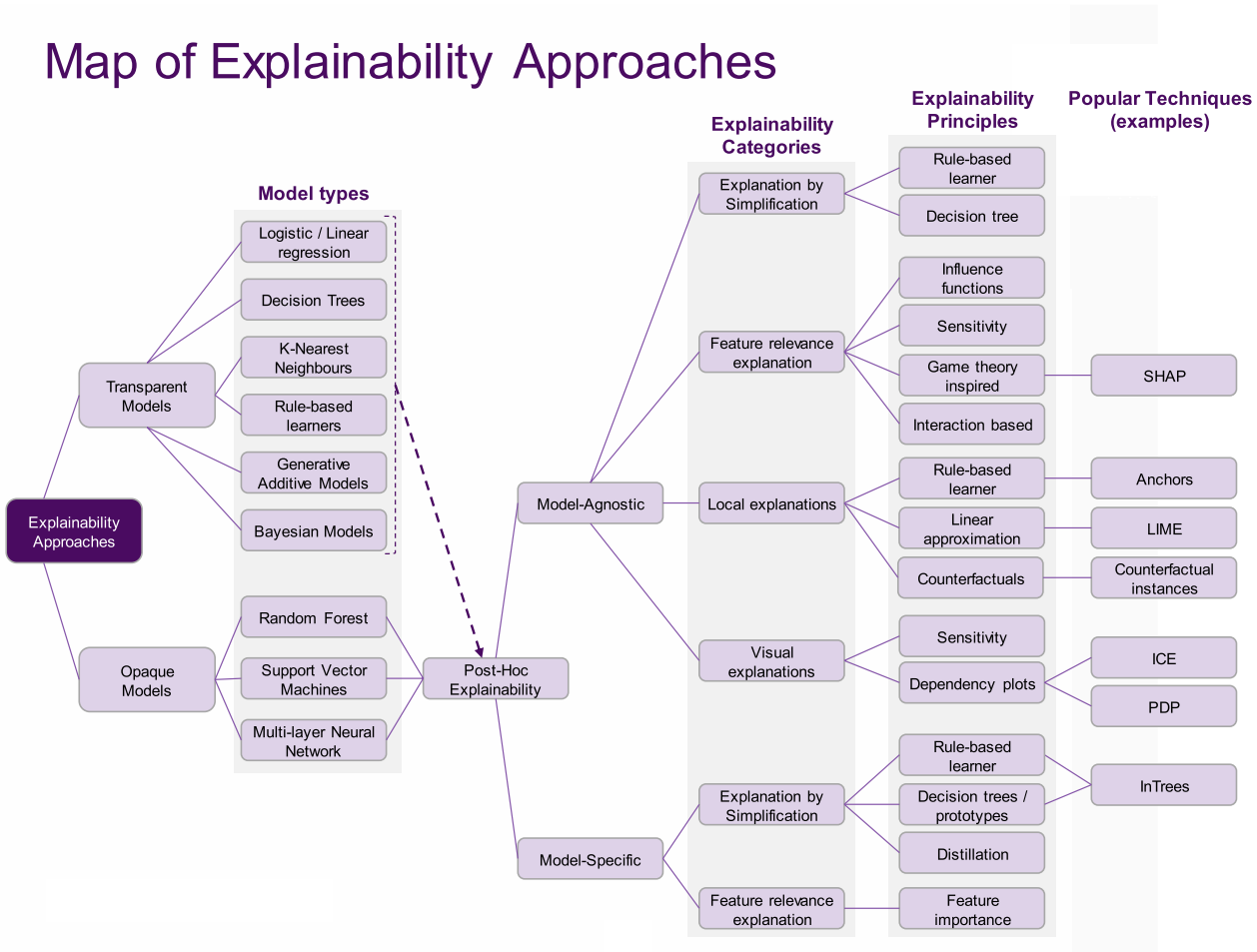


**FIGURE 2 |** A taxonomic view on XAI.

strategy and pipeline we recommend based on the development of numerous libraries for the analysis and interpretation of machine learning models [see, for example (Molnar, 2020)].

## 5.1 Taxonomy Framework

In **Figure 2**, we arrange models in terms of the kinds of explainability that are enabled, to be seen as a taxonomy. The subsequent frameworks will be based on this taxonomy, and can be seen as elaborations on the distinction between transparent and opaque ML models (*Transparency framework*), followed by a description of the capabilities of explainability approaches (*XAI Capability framework*).

## 5.2 Transparency Framework

In **Table 1**, we draw a comparison between models in terms of the kinds of transparency that are enabled. This table demonstrates the correspondence between the design of various transparent ML models and the transparency dimensions they satisfy. Furthermore, it provides a summary of the most common types of explanations that are encountered when dealing with opaque models.

## 5.3 Explainable Machine Learning Capability Framework

In **Table 2**, we draw a comparison between XAI approaches in terms of the type of explanations they offer, whether they are model agnostic and whether they require a transformation of the input data before the method can be applied. This summary can be utilized to distinguish between the capabilities of different explainability approaches, and whether the technical assumptions made for applying the approach (e.g., assumptions about independencies between variables, which is serious and limiting) is a price worth paying for the application at hand.

## 5.4 Explanation Type Framework

In **Table 3**, we contrast the types of post-hoc explanations at a conceptual level: for example, what might local explanations offer in contrast to model simplification strategies?

## 5.5 Data Scientist Strategy Framework

In the penultimate section, we motivate a narrative for a putative data scientist, Jane, and discuss how she might go about

**TABLE 1 |** Comparing models on the kinds of transparency that are enabled.

| Model | Simulatability | Decomposability | Algorithmic transparency | Post-hoc |
|---|---|---|---|---|
| Linear/Logistic regression | Predictors are human readable and interactions among them are kept to a minimum | Too many interactions and predictors | Variables and interactions are too complex to be analyzed without mathematical tools | Not needed |
| Decision trees | Human can understand without mathematical background | Rules do not modify data and are understandable | Humans can understand the prediction model by traversing tree | Not needed |
| K-nearest neighbors | The complexity of the model matches human naive capabilities for simulation | Too many variables, but the similarity measure and the set of variables can be analyzed | Complex similarity measure, too many variables to be analyzed without mathematical tools | Not needed |
| Rule based learners | Readable variables, size of rules is manageable by a human | Size of rules is too large to be analyzed | Rules so complicated that mathematical tools are needed | Not needed |
| General additive models | Variables, interactions and functions must be understandable | Interactions too complex to be simulated | Due to their complexity, variables and interactions cannot be analyzed without mathematical tools | Not needed |
| Bayesian models | Statistical relationships and variables should be understandable by the target audience | Relationships involve too many variables | Relationships and predictors are so complex that mathematical tools are needed | Not needed |
| Tree ensembles | Not applicable | Not applicable | Not applicable | Feature relevance, Model simplification |
| Support vector machines | Not applicable | Not applicable | Not applicable | Feature relevance, Model simplification |
| Multi–layer neural networks | Not applicable | Not applicable | Not applicable | Feature relevance, Model simplification, Visualization |

explaining her models by asking the right questions. We recommend a simple strategy and outline sample questions that motivate certain types of explanations.

In the following sections, we will expand on *transparent models*, followed by *opaque models* and then to *explainability approaches*, all of which are mentioned in the frameworks above.

## 6 TRANSPARENT MODELS

In this section we are going to introduce a set of models that are inherently considered to be transparent. By this, we mean that their intrinsic architecture satisfies at least one of the three transparency dimensions that we discussed in a previous section.

- **Linear\Logistic Regression** refers to a class of models used for predicting continuous\categorical targets, respectively, under the assumption that this target is a linear combination of the predictor variables. That specific modeling choice allows us to view the model as a transparent method. Nonetheless, a decisive factor of how a explainable a model is, has to do with the ability of the user to explain it, even when talking about inherently transparent models. In that regard, although these models satisfy the transparency criteria, they may also benefit from post-hoc explainability approaches (such as visualization), especially when non-expert audience needs to get a better understanding of the models' intrinsic reasoning. The model, nonetheless, has been largely applied within Social Sciences for many decades. As a general remark, we should note that in order for the models to maintain their transparency features, their size must be limited, and the variables used must be understandable by their users.
- **Decision Trees** form a class of models that generally fall into the transparent ML models category. They contain a set of conditional control statements, arranged in a hierarchical manner, where intermediate nodes represent decisions and leaf nodes can be either class labels (for classification problems) or continuous quantities (for regression problems). Supposing a decision tree has only a small amount of features and that its length is not prohibitively long to be memorized by a human, then it clearly falls into the class of simulatable models. In turn, if the model's length does not allow simulating it, but the features are still understandable by a human user, then the model is no longer simulatable, but it becomes decomposable. Finally, if on top of that the model also utilizes complex feature relationships, then it falls into the category of algorithmically transparent models.

Decision trees are usually utilized in cases where understandability is essential for the application at hand, so in these scenarios not overly complex trees are preferred. We should also note that apart from AI and related fields, a significant amount of decision trees' applications come from other fields, such as medicine. However, a major limitation of these models stems from their tendency to overfit the data, leading to poor generalization performance, hindering their application in cases where high predictive accuracy is desired. In such cases, ensembles of trees could offer much better generalization, but these models cannot be considered transparent anymore[4].

[4]Although an ensemble of a small number of decision trees could still fall under the category of transparent models, those employed in real-world applications typically consist of a large number of trees so can be seen to lose transparency properties.

**TABLE 2 |** Comparing XAI methods.

| XAI method | Swapping | Explanation | Model agnostic | Categorical/ Continuous features | Intermediate transformation | Independent features | Shapley values | Examples |
|---|---|---|---|---|---|---|---|---|
| KernelSHAP (Lundberg and Lee, 2017) | No | Feature relevance | Yes | Both | Yes | Yes | Yes | No |
| TreeSHAP (Lundberg and Lee, 2017) | Yes | Feature relevance | No | Both | Yes | No | Yes | No |
| LIME (Ribeiro et al., 2016) | Yes | Simplification | Yes | Both | Yes | No | Not necessarily | No |
| Anchors (Ribeiro et al., 2018) | Yes | Simplification | Yes | Both | No | No | No | No |
| QII (Datta et al., 2016) | Yes | Feature relevance | Yes | Both | Yes | No | Not necessarily | No |
| CNF rules (Su et al., 2016) | Yes | Simplification | Yes | Categorical | No | No | No | No |
| Influence function (Koh and Liang, 2017) | Yes | Feature relevance | Yes | Both | No | No | No | Yes |
| ASTRID (Henelius et al., 2017) | Yes | Feature relevance | Yes | Both | No | No | No | No |
| Distilation (Tan et al., 2017) | Yes | Simplification | Yes | Both | No | No | No | No |
| Counterfactual (Wachter et al., 2018) | Yes | Local | Yes | Both | No | No | No | Yes |
| InTrees (Deng, 2014) | Yes | Simplification | No | Both | No | No | No | No |
| Prototypes (Tan et al., 2016) | Yes | Simplification | No | Both | No | No | No | Yes |
| Feature tweaking (Tolomei et al., 2017) | Yes | Feature relevance | No | Both | No | No | No | Yes |

**TABLE 3 |** Advantages and disadvantages of the various kinds of explanations.

| Explanation | Advantages | Disadvantages |
|---|---|---|
| Local explanations | Explains the model's behaviour in a local area of interest. Operates on instance-level explanations. | Explanations do not generalize on a global scale. Small perturbations might result in very different explanations. Not easy to define locality. Some approaches face stability issues. |
| Examples | Representative examples provide insights about the model's internal reasoning. Some of the algorithms uncover the most influential training data points that led the model to its predictions. | Examples require human inspection. They do not explicitly state what parts of the example influence the model. |
| Feature relevance | They operate on an instance level, calculating the importance of each feature in the model's decision. A number of the proposed approaches come with appealing theoretical guarantees. | They are sensitive in cases where the features are highly correlated. In many cases the exact solutions are approximated, leading to undesirable side effects, such as the ordering affecting the outcome. |
| Simplification | Simple surrogate models explain the opaque ones. Resulting explanations, such as rules, are easy to understand. | Surrogate models may not approximate the original models well. Surrogate models come with their own limitations. |
| Visualizations | Easier to communicate to non technical audience. Most of the approaches are intuitive and not hard to implement. | There is an upper bound on how many features we can consider at once. Humans need to inspect the resulting plots in order to produce explanations. |

- **K-Nearest Neighbors (KNN)** is also a method that falls within transparent models, which deals with classification problems in a simple and straightforward way: it predicts the class of a new data point by inspecting the classes of its K nearest neighbors (where the neighborhood relation is induced by a measure of distance between data points). The majority class is then assigned to the instance at hand.

Under the right conditions, a KNN model is capable of satisfying any level of transparency. It should be noted, however, that this depends heavily on the distance function that is employed, as well as the model's size and the features' complexity, as in all the previous cases.

- **Rule-based learning** is built on the intuitive basis of producing rules to describe how a model generates its outputs. The complexity of the resulting rules ranges from simple "if-else" expressions to fuzzy rules, or propositional rules encoding complex relationships between variables. As humans also utilize rules in everyday life, these systems are usually easy to understand, meaning they fall into the category of transparent models. Having said that, the exact level of transparency depends on some designing aspects, such as the coverage (amount) and the specificity (length) of the generated rules.

Both of these factors are at odds with the transparency of the resulting model. For example, it is reasonable to expect that a

system with a very large amount of rules is infeasible to be simulated by a human. The same applies to rules containing a prohibiting number of antecedents or consequents. Including cumbersome features in the rules, on top of that, could further impede their interpretability, rendering system just algorithmically transparent.

- **Generalized Additive Models (GAMs)** are a class of linear models where the outcome is a linear combination of some functions of the input features. The goal of these models is to infer the form of these unknown functions, which may belong to a parametric family, such as polynomials, or they could be defined non-parametrically. This allows for a large degree of flexibility, since at some applications they may take the form of a simple function, or be handcrafted to represent background knowledge, while in others they may be specified by just some properties, such as being smooth.

These models certainly satisfy the requirements for being algorithmic transparent, at least. Furthermore, in applications where the dimensionality of the problem is small and the functions are relatively simple, they could also be considered simulatable. However, we should note that while utilizing non-parametric functional forms may enhance the models fit, it comes with a trade-off regarding its interpretability. It is also worth noting that, as with linear regression, visualization tools are often employed to communicate the results of the analysis [such as partial dependence plots (Friedman and Meulman, 2003)].

- **Bayesian networks** refer to the designing approach where the probabilistic relationships between variables are explicitly represented using a directed graph, usually an acyclic one. Due to this clear characterization of the connection among the variables, as well as graphical criteria that examine probabilistic relationships by only inspecting the graphs topology (Geiger et al., 1990), they have been used extensively in a wide range of applications (Kenett, 2012; Agrahari et al., 2018).

Following the above, it is clear that they fall into the class of transparent model. They can potentially fulfill the necessary prerequisites to be members of all three transparency levels, however including overly complex features or complicating graph topologies can result into them satisfying just algorithmic transparency. Research into model abstractions may be relevant to address this issue (John, 2017; Belle, 2019).

Owing to their probabilistic semantics, which allows conditioning and interventions, researchers have looked into ways to augment directed and undirected graphical models (Baum and Petrie, 1966) further to provide explanations, although, of course, they are already inherently transparent in the sense described above. Relevant works include (Timmer et al., 2016), where the authors propose a way to construct explanatory arguments from Bayesian models, as well as (Kyrimi et al., 2020), where explanations are produced in order to assess the trustworthiness of a model. Furthermore, ways to draw representative examples from data have been considered, such as in (Kim et al., 2014).

A general remark, even when utilizing the models discussed above, is about the trade-off between complexity and transparency. Transparency, as a property, is not sufficient to guarantee that a model will be readily explainable. As we saw in the above paragraphs, as certain aspects of a model become more complex, it is not apparent how it operates internally, anymore. In these cases, XAI approaches could be used to explain the model's decisions, while utilizing an opaque model could also be considered.

## 7 OPAQUE MODELS

While the models we discussed in the previous section come with appealing transparency features, it is not always that they are among the better performing ones, at least as determined by predictive accuracy on standard (say) vision datasets. In this section we will touch on the class of opaque models, a set of ML models which, at the expense of explainability, achieve higher accuracy utilizing complex decision boundaries.

- **Random Forests (RF)** were initially proposed as a way to improve the accuracy of single decision trees, which in many cases suffer from overfitting, and consequently, poor generalization. Random forests address this issue by combining multiple trees together, in an attempt to reduce the variance of the resulting model, leading to better generalization (Hastie et al., 2008). In order to achieve this, each individual tree is trained on a different part of the training dataset, capturing different characteristics of the data distribution, to obtain an aggregated prediction. This procedure results in very expressive and accurate models, but it comes at the expense of interpretability, since the whole forest is far more challenging to explain, compared to single trees, forcing the user to apply post-hoc explainability techniques to gain an understanding of the decision machinery.
- **Support Vector Machines (SVMs)** form a class of models rooted deeply in geometrical approaches. Initially introduced for linear classification (Vapnik and Lerner, 1963), they were later extended to the non-linear case (Boser et al., 1992), while a relaxation of the original problem (Cortes and Vapnik, 1995) made it suitable for real-life applications. Intuitively, in a binary classification setting, SVMs find the data separating hyperplane with the maxim margin, meaning the distance between it and the nearest data point of each class is as large as possible. Apart from classification purposes, SVMs can be applied in regression (Drucker et al., 1996), or even clustering problems (Ben-Hur et al., 2001). While SVMs have been successfully used in a wide array of applications, their high dimensionality as well as potential data transformations and

geometric motivation, make them very complex and opaque models.

- **Multi-layer Neural Networks (NNs)** are a class of models that have been used extensively in a number of applications, ranging from bioinformatics (Chicco et al., 2014) to recommendation systems (van den Oord et al., 2013), due to their state-of-the-art performance. On the other hand, their complex topology hinders their interpretability, since it is not clear how the variables interact with each other or what kind of high level features the network might has picked up. Furthermore, even the theoretical/mathematical understanding of their properties has not been sufficiently developed, rendering them virtual black-box models.

From a technical point of view, NNs are comprised of successive layers of nodes connecting the input features to the target variable. Each node in an intermediate layer collects and aggregates the outputs of the preceding layer and then produces an output on its own, by passing the aggregated value through a function (called activation function).[5] In turn, these values are passed on to the next layer and this process is continued until the output layer is reached.

An immediate observation is that as the number of layers increases, the harder it becomes to interpret the model. In contrast, an overly simple NN could even fall into the class of simulatable models. But such a simple model is of very little practical interest these days.

# 8 EXPLAINABILITY APPROACHES

In this section, we are going to review the literature and provide an overview of the various methods that have been proposed in order to produce post-hoc explanations from opaque models. The rest of the section is divided into the techniques that are especially designed for Random Forests and then we turn to ones that are model agnostic. We focus on Random Forests owing to their popularity and to illustrate an emerging literature on model-specific explainability which often leverages technical properties of the ML model to provide a more sophisticated or otherwise customized explainability approach.

## 8.1 Random Forest Explainability Approaches

As discussed above, Random Forests are among the best performing ML algorithms, used in a wide variety of domains. However, their performance comes at the cost of explainability, so bespoke post-hoc approaches have been developed to facilitate the understanding of this class of models. For tree ensembles, in general, most of the techniques found in the literature fall into either the explanation by simplification or feature relevance explanation categories. In what follows, we review some of the most popular approaches.

### 8.1.1 Simplifying and Extracting Rules

An attempt to simplify RFs can be found in (Hara and Hayashi, 2016), where the authors propose a way to approximate them with a mixture of simpler models. The only requirement for the user is to specify the number of rules that the new mixture of models should contain, thereby providing a degree of freedom regarding how many rules are required to distill the model's intrinsic reasoning. Then, the resulting mixture approximates the original model utilizing only the amount of rules that the user specified.

Other approaches, similar in spirit, can be found in (Van Assche and Blockeel, 2007; Zhou and Hooker, 2016), where the objective is to approximate the RF using a single decision tree. In (Van Assche and Blockeel, 2007), the authors utilize a heuristic, based on information gain, in order to construct a tree that is compact enough to retain interpretability. On the other hand, the approach in (Zhou and Hooker, 2016) was based on studying the asymptotic behavior of the Gini index, in order to train a stable and accurate decision tree.

Another approach to simplify RFs is discussed in (Deng, 2014). The main contribution of this work is proposing a methodology for extracting the more representative rules a RF has acquired. This approach remedies the fact that RFs consist of thousands of rules: by selecting only the most prominent ones, the amount is reduced drastically. In this case, too, the resulting rules approximate the original model, but the difference is that now rules are not learnt by a new model, but are extracted from the RF directly. Furthermore, the obtained rules can easily be combined in order to create a new rule based classifier.

The idea above, has been explored from other perspectives as well. In (Mashayekhi and Gras, 2015), a different method for extracting rules from a RF is proposed. In this case, a hill climbing methodology is utilized in order to construct a set of rules that approximates the original RF. This, again, leads to a significantly smaller set of rules, facilitating the model's comprehensibility.

A line of research that has also been explored for producing explanations when using RFs is by providing the user with representative examples. The authors in (Tan et al., 2016) examine ways to inspect the training dataset in order to sample a number of data points that are representative members of their corresponding class. This method comes with some theoretical guarantees about the quality of the examples, while it is also adaptive, in the sense that the user specifies the number of total examples, and then the algorithm decides how many examples to sample from each class.

### 8.1.2 Feature Relevance

Along with simplification procedures, feature relevance techniques are commonly used for tree ensembles. One of the first approaches can be found in Breiman (Breiman et al., 1984). His method is based on permuting the values of a feature within the dataset, and then utilizing various metrics to assess the

[5]It is worth noting that there are a number of options when it comes to specifying the activation function, which along with the number of the intermediate layers determine the quality of the resulting model.

difference in quality between the original and the newly acquired model.

The authors of (Palczewska et al., 2013) develop an approach for assessing the importance of individual features, by computing how much the model's accuracy drops, after excluding a feature. Furthermore, employing this method it is possible to extract a prototypical vector of feature contributions, so we can get an idea of how important each feature is, with respect to the instances belonging in a given class. It is worth noting that extensions of this method in service of communicating explainability have been proposed as well, such as in (Welling et al., 2016), where, in addition to slightly modifying the way a feature's importance is computed, graphical tools for visualizing the results are developed.

A different approach on measuring a feature's importance can be found in (Tolomei et al., 2017). The aim of this work is to examine ways to produce "counterfactual" data points, in the following sense: assuming a data point was classified as negative (positive), how can we generate a new data point, as similar as possible to the original one, that the model would classify as positive (negative)? The similarity metric is given by the user, so it can be application specific, incorporating expert knowledge. A by-product of this procedure is that by examining the extent to which a feature was modified, we get an estimate of its importance, as well as the new counterfactual data point.

In a somewhat different, yet relevant, approach the authors in (Petkovic et al., 2018) develop a series of metrics assessing the importance of the model's features. Apart from standard importance scores, they also discuss how to answer more complex questions, such as what is the effect on the model's accuracy, when using only a subset of the original features, or which subsets of features interact together.

Other ways to identify a set of important features can be found in the literature, as well. The authors in (Auret and Aldrich, 2012) propose a way to determine a threshold for identifying important features. All features exceeding this threshold are deemed important, while those that do not are discarded as unnecessary. Following this approach, apart from having a vector with each feature's importance, a way to identify the irrelevant ones is also provided. In addition, graphical tools to communicate the results to a non-expert audience are discussed.

## 8.2 Model-Agnostic Explainability Approaches

Model-agnostic techniques are designed having the purpose of being generally applicable, in mind. They have to be flexible enough, so they do not depend on the intrinsic architecture of a model, thus operating solely on the basis of relating the input of a model to its outputs. Arguably, the most prominent explanation types in this class are model simplification, feature relevance, as well as visualizations.

### 8.2.1 Explanation by Simplification

Arguably the most popular is the technique of Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016). LIME approximates an opaque model locally, in the surrounding area of the prediction we are interested in explaining, building either a linear model or a decision tree around the predictions of an opaque model, using the resulting model as a surrogate in order to explain the more complex one. Furthermore, this approach requires a transformation of the input data to an "interpretable representation," so the resulting features are understandable to humans, regardless of the actual features used by the model (this is termed as "intermediate transformation," in **Table 2**).

A similar technique, called anchors, can be found in (Ribeiro et al., 2018). Here the objective is again to approximate a model locally, but this time not by using a linear model. Instead, easy to understand "if-then" rules that anchor the model's decision are employed. The rules aim at capturing the essential features, omitting the rest, so it results in more sparse explanations.

G-REX (Konig et al., 2008) is an approach first introduced in genetic programming, in order to extract rules from data, but later works have expanding its score, rendering capable of addressing explainability (Johansson et al., 2004a; Johansson et al., 2004b).

Another approach is introduced in (Su et al., 2016), where the authors explore a way to learn rules in either Conjunctive Normal Form (CNF) or Disjunctive Normal Form (DNF). Supposing that all variables are binary, then the algorithm builds a classification model that attempts to explain the complex model's decisions utilizing only such propositional rules. Such approaches have the extra benefit of resulting in a set of symbolic rules that are explainable by default, as well as can be utilized as a predictive model, themselves.

Another perspective in simplification is introduced in (Krishnan and Wu, 2017). In this work, the objective is to approximate an opaque model using a decision tree, but the novelty of the approach lies on partitioning the training dataset in similar instances, first. Following this procedure, each time a new data point is inspected, the tree responsible for explaining similar instances will be utilized, resulting in better local performance. Additional techniques to construct rules explaining a model's decisions can be found in (Turner, 2016a; Turner, 2016b).

In similar spirit, the authors of (Bastani et al., 2017) formulate model simplification as a model extraction process by approximating a complex model using a transparent one. The proposed approach utilizes the predictions of a black-box model to build a (greedy) decision tree, in order to inspect this surrogate model to gain some insights about the original one. Simplification is approached from a different perspective in (Tan et al., 2017), where an approach to distill and audit black box models is presented. This is a two-part process, comprising of a distillation approach, as well as a statistical test. So, overall, the approach provides a way to inspect whether a set of variables is enough to recreate the original model, or if extra information is required in order to achieve the same accuracy.

There has been considerable recent development in the so-called counterfactual explanations (Wachter et al., 2018). Here, the objective is to create instances as close as possible to the instance we wish to explain, but such that the model classifies the new instance in a different category. By inspecting this new

data point and comparing it to the original one we can gain insights on what the model considers as minimal changes to the original data point, so as to change its decision. A simple example is the case of an applicant who was denied his loan application, and the explanation might say that had he had a permanent contract with his current employer, the loan would be approved.

### 8.2.2 Feature Relevance

One of the most popular contributions here, and in XAI in general, is that of SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017). The objective in this case is to build a linear model around the instance to be explained, and then interpret the coefficients as the feature's importance. This idea is similar to LIME, in fact LIME and SHAP are closely related, but SHAP comes with a set of nice theoretical properties. Its mathematical basis is rooted in coalitional Game Theory, specifically on Shapley values (Shapley, 1952). Roughly, the Shapley value of a feature is its average expected marginal contribution to the model's decision, after all possible combinations have been considered. However, the dimensionality of many complex real-world applications renders the calculation of these values infeasible, so the authors in (Lundberg and Lee, 2017) simplify the problem by making various assumptions, such as independency among the variables. Arguably, this is a strong assumption that can affect the quality of the resulting Shapley values. Other issues exist as well, for example while in its formulation all possible orderings of the variables are considered, in practice this is infeasible, so the ordering of the variable affects the computation of the Shapley values (In **Table 2**, for example, we use the term "swapping" to refer to whether a method is influenced by the features' ordering).

Similar in spirit, in (Strumbelj and Kononenko, 2010), the authors propose to measure a feature's importance using its Shapley value, but the objective function, as well as the optimization approach, is not the same as in SHAP. A different strategy is considered in (Datta et al., 2016), where a broad variety of measures are presented to tackle the quantification of the degree of influence of inputs on the outputs. The proposed QII (Quantitative Input Influence) measures account for correlated inputs, which quantifies the influence by estimating the change in performance when using the original data set vs. when using one where the feature of interest is replaced by a random quantity.

In relation to the above, it is worth mentioning that the concept of Shapley values has proven to be highly influential within the XAI community. On the one hand, the popularity of SHAP naturally led to further research, aiming to design complimentary tools to better understand its outcomes. For example (Kumar I. et al., 2020), presents a diagnostic tool that can be useful when interpreting Shapley values, since these scores alone can be misleading, as the authors argue. Another interesting development can be found in (Joseph, 2019), where a series of statistical tests are developed, allowing for producing confidence intervals for the resulting Shapley values. Furthermore, such approaches are also significant since they draw connections between well-known statistical techniques and XAI, expanding the range of the latter, while also opening the door for utilizing tools that could possibly address current robustness issues.

On the other hand, research has also looked into connecting Shapley values and statistics in alternative ways as well. A representative example can be found in (Song and Barry, 2016; Owen and Prieur, 2017), where the authors demonstrate how Shapley values can be used to quantify variable importance, instead of functional ANOVA (Owen, 2013), which decomposes a function into orthogonal components and defines importance measures based on them. This is shown to be particularly powerful when there is dependence between the variables, alleviating a series of limitations of existing techniques (Chastaing et al., 2012). Another recent development can be found in (Giudici and Raffinetti,), where the authors combine Lorenz Zonoids (Koshevoy and Mosler, 1996), a generalization of ROC curves (Fawcett, 2006), with Shapley values. The result is a technique that combines local attributions with predictive accuracy, in a manner that is simple and relatively easy to interpret, since it connects to various well studied statistical measures.

Another approach that is based on random feature permutations can be found in (Henelius et al., 2014). In this work, a methodology for randomizing the values of a feature, or a group of features, is introduced, based on the difference between the model's behavior when making predictions for the original dataset and when it does the same for the randomized version. This process facilitates the identification of important variables or variable interactions the model has picked up.

Additional ways to assess the importance of a feature can also be found, such as the one in (Adebayo and Kagal, 2016). The authors introduce a methodology for computing feature importance, by transforming each feature in a dataset, so the result is a new dataset where the influence of a certain feature has been removed, meaning that the rest of the attributes are orthogonal to it. By using several modified datasets, the authors develop a measure for calculating a score, based on the difference in the model's performance across the various datasets.

Different from the above threads, in (Cortez and Embrechts, 2011), the authors extend existing SA (Sensitivity Analysis) approaches in order to design a Global SA method. The proposed methodology is also paired with visualization tools to facilitate communicating the results. Likewise, the work in (Henelius et al., 2017) presents a method (ASTRID) that aims at identifying which attributes are utilized by a classifier in prediction time. They approach this problem by looking for the largest subset of the original features so that if the model is trained on this subset, omitting the rest of the features, the resulting model would perform as well as the original one. In (Koh and Liang, 2017), the authors use influence functions to trace a model's prediction back to the training data, by only requiring an oracle version of the model with access to gradients and Hessian-vector products. Finally, another way to measure a data point's influence on the model's decision comes from deletion diagnostics (Cook, 1977). The difference this time is that this approach is concerned with measuring how omitting a

data point from the training dataset influences the quality of the resulting model, making it useful for various tasks, such as model debugging.

### 8.2.3 Visual Explanations

Some popular approaches to visualizations can be found in (Cortez and Embrechts, 2011), where an array of various plots are presented. Additional techniques are discussed in (Cortez and Embrechts, 2013), where some new SA approaches are introduced. Finally (Friedman, 2001; Goldstein et al., 2013), presents ICE (Individual Conditional Expectation) and PD (Partial Dependence) plots, respectively. The former, operates on instance level, depicting the model's decision boundary as a function of a single feature, with the rest of them staying fixed to their observed values. This way it is possible to inspect that feature's effect on the model's decisions, under the specific context that is formed by the remaining variables. In contrast, the latter plots the model's decision boundary as a function of a single feature when the remaining features are averaged out, so this shows the average effect of that feature to the model's outcome. PDPs provide insights about the form of the relationship between the feature of interest and the outcome, such as whether it is linear, monotonic, or more complex (Molnar, 2020). On the other hand, average effects can be potentially misleading, hindering the identification of interactions among the variables. In turn, a more complete approach would be to utilize both plots, due to their complementary nature. This is also enforced by observing there is an interesting relationship between these two plots, as averaging the ICE plots of each instance of a dataset, yields the corresponding PD plot.

Along with the three frameworks, the above exposition covers the main observations and properties of XAI trends.

## 9 BRIEF OVERVIEW OF DEEP LEARNING MODELS

In this section we provide a brief summary of XAI approaches that have been developed for deep learning (DL) models, specifically multi-layer neural networks (NNs). NNs are highly expressive computational models, achieving state-of-the-art performance in a wide range of applications. Unfortunately, their architecture and learning regime corresponds to a complex computational pipeline, so they do not satisfy any level of transparency, at least when we go beyond simple models, such as single layer perceptron as mentioned previously, although, of course such models do not fall within "deep" learning. This has led to the development of NN-specific XAI methods, utilizing their specific topology. The majority of these methods fall into the category of either *model simplification* or *feature relevance*.

In *model simplification*, rule extraction is one of the most prominent approaches. Rule extraction techniques that operate on a neuron-level rather than the whole model are called decompositional (Özbakundefinedr et al., 2010). proposes a method for producing if-else rules from NNs, where model training and rule generation happen at the same time. CRED (Sato and Tsukimoto, 2001) is a different approach that utilizes decision trees to represent the extracted rules. KT (Fu, 1994) is a related algorithm producing if-else rules, in a layer by layer manner. DeepRED (Zilke et al., 2016) is one of the most popular such techniques, extending CRED. The proposed algorithm has additional decision trees as well as intermediate rules for every hidden layer. It can be seen as a divide and conquer method aiming at describing each layer by the previous one, aggregating all the results in order to explain the whole network.

On the other hand, when the internal structure of a NN is not taken into account, the corresponding methods are called pedagogical. That is, approaches that treat the whole network as a black-box function and do not inspect it at a neuron-level in order to explain it. TREPAN (Craven and Shavlik, 1994) is such an approach, utilizing decision trees as well as a query and sample approach. Saad and Wunsch (Saad and Wunsch, 2007) have proposed an algorithm called HYPINV, based on a network inversion technique. This algorithm is capable of producing rules having the form of the conjunction and disjunction of hyperplanes. Augusta and Kathirvalavakumar (Augasta and Kathirvalavakumar, 2012) have introduced the RxREN algorithm, employing reverse engineering techniques to analyze the output and trace back the components that cause the final result.

Combining the above approaches leads to eclectic rule extraction techniques. RX (Hruschka and Ebecken, 2006) is such a method, based on clustering the hidden units of a NN and extracting logical rules connecting the input to the resulting clusters. An analogous eclectic approach can be found in (Kahramanli and Allahverdi, 2009), where the goal is to generate rules from a NN, using so-called artificial immune system (AIS) (Dasgupta, 1999) algorithms.

Apart from rule extraction techniques, other approaches have been proposed to interpret the decisions of NNs. In (Che et al., 2016), the authors introduce *Interpretable Mimic Learning*, which builds on model distillation ideas, in order to approximate the original NN with a simpler, interpretable model. The idea of transferring knowledge from a complex model (the *teacher*) to a simpler one (the *student*) been explored in other works, for example (Bucila et al., 2006; Hinton et al., 2015; Micaelli and Storkey, 2019).

An intuitive observation about NNs is that as the number of layers grows larger, developing model simplification algorithms gets progressively more difficult. Due to this, *feature relevance* techniques have gained popularity in recent years. In (Kindermans et al., 2017), the authors propose ways to estimate neuron-wize signals in NNs. Utilizing these estimators they present an approach to superposition neuron-wize explanations in order to produce more comprehensive explanations.

In (Montavon et al., 2017) a way to decompose the prediction of a NN is presented. To this end, a neuron's activation is decomposed and then its score is backpropagated to the input layer, resulting in a vector containing each feature's importance.

DeepLIFT (Shrikumar et al., 2017) is another way to assign importance scores when using NNs. The idea behind this method

is to compare a neuron's activation to a reference one and then use their difference to compute the importance of a feature.

Another popular approach can be found in (Sundararajan et al., 2017), where the authors present Integrated Gradients. In this work, the main idea is to examine the model's behavior when moving along a line connecting the instance to be explained with a baseline instance (serving the purpose of a "neutral" instance). Furthermore, this method comes with some nice theoretical properties, such as *completeness* and *symmetry preservation*, that provide assurances about the generated explanations.

## 10 VIEWS AND SUGGESTIONS

XAI is a broad and relatively new branch of ML, which, in turn, means that there is still some ambiguity regarding the goals of the resulting approaches. The approaches presented in this survey are indicative of the range of the various explainability angles that are considered within the field. For example, feature relevance approaches provide insights by measuring and quantitatively ranking the importance of a feature, model simplification approaches construct relatively simple models as proxies for the opaque ones, while visual explanations inspect a model's inner understanding of a problem through graphical tools. At this point we should note that choosing the right technique for the application at hand depends exactly at the kind of insights the user would like to gain, or perhaps the kind of explanations he/she is more comfortable interpreting.

In applications where explainability is of utmost importance, it is worth considering using a transparent model. The downside of this, is that these models often compromise performance for the sake of explainability, so it is possible that the resulting accuracy hinders their employment in crucial real-world applications.

In cases where maintaining high accuracy is a driving factor, too, employing an opaque model and pairing it with some XAI techniques, instead of using a transparent one, is probably the most reasonable choice. Subsequently, identifying the right technique for explaining the resulting model is the next step in the quest to understand its internal mechanisms. Each of them comes with its own strong points, as well as limitations. More specifically:

- Local explanations approximate the model in a narrow area, around a specific instance of interest. They offer information about how the model operates when encountering inputs that are similar to the one we are interested in explaining. This information can attain various forms, such as importance scores or rules. Of course, this means that the resulting explanations do not necessarily reflect the model's mechanism on a global scale. Other limitations arise when considering the inherent difficulty to define what a local area means in a high dimensional space. This could also lead to cases where slightly perturbing a feature's value results in significantly different explanations.
- Representative examples allow the user to inspect how the model perceives the elements belonging in a certain category. In a sense, they serve as prototype data points. In other related approaches, it is possible to trace the model's decision back to the training dataset and uncover the instance that influenced the model's decision the most. Deletion diagnostics also fall into this category, quantifying how the decision boundary changes when some training datapoints are left out. The downside of utilizing examples is that they require human inspection in order to identify the parts of the example that distinguish it from the other categories.
- Feature relevance explanations aim at computing the influence of a feature in the model's outcome. This could be seen as an indirect way to produce explanations, since they only indicate a feature's individual contribution, without providing information about feature interactions. Naturally, in cases where there are strong correlations among features, it is possible that the resulting scores are counterintuitive. On the other hand, some of these approaches, such as SHAP, come with some nice theoretical properties [although in practice they might be violated (Merrick and Taly, 2019; Kumar I. E. et al., 2020)].
- Model simplification comes with the immediate advantage and flexibility of allowing to approximate an opaque model using a simpler one. This offers a wide range of representations that can be utilized, from simple "if-then" rules to fitting surrogate models. This way explanations can be adjusted to best fit a particular audience. Of course, there are limitations as well, with perhaps the most notable one being the quality of the approximation. Furthermore, usually, it is not possible to quantitatively assess it, so empirical demonstrations are needed to illustrate the goodness of the approximation.
- Visualizations provide for a way to utilize graphical tools to inspect some aspects of a model, such as its decision boundary. In most cases they are relatively easy to understand for both technical and non technical audiences. However, when resorting to visualizations, many of the proposed approaches make assumptions about the data (such as independence) that might not hold for the particular application, perhaps distorting the results.

Overall, we summarize some of the salient properties to consider in **Table 3**.

Taking a close look at the various kinds of explanations discussed above, makes clear that each of them addresses a different aspect of explainability. This means that there is no approach suitable for each and every scenario. This is in tune with how humans perceive explainability as well, since we know that there is not a single question whose answer would be able to communicate all the information needed to explain any situation. Most of the times, one would have to ask multiple questions, each one dealing with a different aspect of the situation in order to obtain a satisfactory explanation.

The same approach should be utilized when inspecting the reasoning of ML models. Relying on only one technique will only give us a partial picture of the whole story, possibly missing out important information. Hence, combining multiple approaches together provides for a more cautious way to explain a model.

At this point we would like to note that there is no established way of combining techniques (in a pipeline fashion), so there is

room for experimenting and adjusting them, according to the application at hand. Having said that, we think that a reasonable base case could look like this:

- If explainability is essential for the application, first try transparent models.
- If it doesn't perform well, and particularly if the complexity of the model is escalating and you lose the explainability benefit, use an opaque one.
- Employ an feature relevance method to get the an instance-specific estimate of each feature's influence.
- A model simplification approach could be used to inspect whether the important features, will turn out to be important on a global scale, too.
- A local explanation approach could shed light into how small perturbations affect the model's outcome, so pairing that with the importance scores could facilitate the understanding of a feature's significance.
- A visualization technique to plot the decision boundary as a function of a subset of the important features, so we can get a sense of how the model's predictions change.

## 11 JANE, THE DATA SCIENTIST

In this section we will discuss a concrete example of how a data scientist could apply the insights gained so far, in a real-life scenario. To this end, we would like to introduce Jane, a data scientist whose work is on building ML models for loan approvals. As a result, she would like to consider things like the likelihood of default given some parameters in a credit decision model.

Jane's current project is to employ a model to assess whether a loan should be approved, that maximizes performance while also maintaining explainability (cf. **Figure 3**).[6] This leads to the challenge of achieving an appropriate trade-off between these two things. Broadly, we can think of two possible choices for Jane (cf. **Figure 4**):

- She can go for transparent models, resulting in a clear interpretation of the decision boundary, allowing for immediately interpreting how a decision is made. For example, if using logistic regression, the notion of defaulting can seen as a weighted sum of features, so a feature's coefficient will tell you this feature's impact on predicting a loan default.
- Otherwise, she can go for an opaque model, which usually achieves better performance and generalizability than its transparent counterparts. Of course, the downside is that in this case is it will not be easy to interpret the model's decisions.

[6]Note that this informal view encourages a notional plot of explainability versus accuracy, as is common in informal discussions on the challenge of XAI (Gunning, 2017; Weld and Bansal, 2019). However, this informal view has been criticized (Rudin, 2019) as being misleading. Since we are concerned primarily with mainstream ML models and the interpretability that emerges when applying statistical analysis to such models, we will continue using this notional idea for the sake of simplicity.

Jane decides to give various transparent models a try, but the resulting accuracy is not satisfactory, so she resorts to opaque models. She again tries various candidates and she finds out that Random Forests achieve the best performance among them, so this is what she will use. The downside is that the resulting model is not immediate to explain anymore (cf. **Figure 5**). In turn, after training the model, the next step is to come up with ways that could help her explain how the model operates to the stakeholders.

The first thing that came to Jane's mind was to utilize one of the most popular XAI techniques, SHAP. She goes on applying it to explain a specific decision made by the model. She computes the importance of each feature and shares it with the stakeholders to help them understand how the model operates. However, as the discussion progresses, a reasonable question comes up (**Figure 6**): could it be that the model relies heavily on an applicant's salary, for example, missing other important factors? How would the model perform on instances where applicants have a relatively low salary? For example, assuming that everything else in the current application was held intact, what is the salary's threshold that differentiates an approved from a rejected application?

These questions cannot been addressed using SHAP, since they refer to how the model's predictive behavior would change, where SHAP can only explain the instance at hand, so Jane realizes that she will have to use additional techniques to answer these questions. To this end, she decides to employ Individual Conditional Expectation (ICE) plots, to inspect the model's behavior for a specific instance, where everything except salary is held constant, fixed to their observed values, while salary is free to attain different values. She could also compliment this technique using Partial Dependence Plots (PDPs) to plot the model's decision boundary as a function of the salary, when the rest of the features are averaged out. This plot allows her to gain some insights about the model's average behavior, as the salary changes (**Figure 7**).

Jane discusses her new results with the stakeholders, explaining how these plots provide answers to the questions that were raised, but this time there is a new issue to address. In the test set there is an application that the model rejects, which comes contrary to what various experts in the bank think should have happened. This leaves the stakeholders in question of why the model decides like that and whether a slightly different application would have been approved by the model. Jane decides to tackle this using counterfactuals, which inherently convey a notion of "closeness" to the actual world. She applies this approach and she finds out that it was the fact that the applicant had missed one payment that led to this outcome, and that had he/she missed none the application would had been accepted (**Figure 8**).

The stakeholders think this is a reasonable answer, but now that they saw how influential the number of missed payments was, they feel that it would be nice to be able to extract some kind of information explaining how the model operates for instances that are similar to the one under consideration, for future reference.

Jane thinks about it and she decides to use anchors in order to achieve just that, generate easy-to-understand "if-then" rules that approximate the opaque model's behavior in a local area (**Figure 9**). The resulting rules would now look something like "if salary is greater than 20 $k$ £ and there are no missed payment, then the loan is approved."

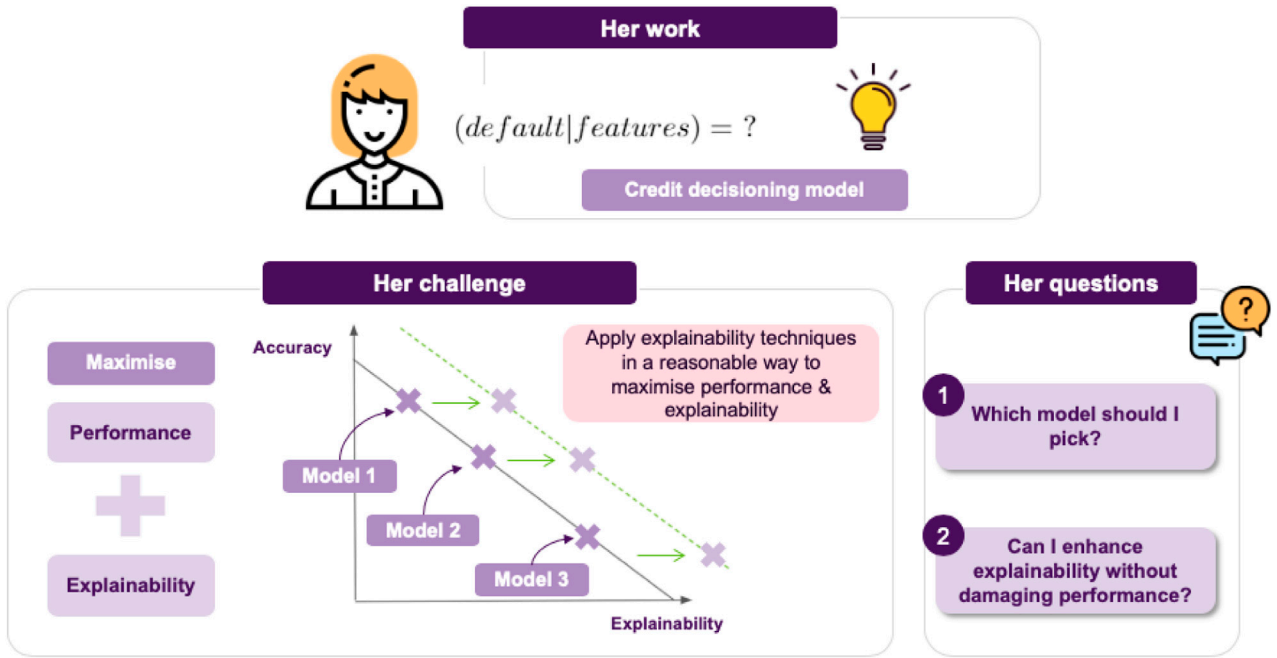


**FIGURE 3 |** Jane's agenda and challenge: which model offers the best trade-off in terms of accuracy vs. explainability?

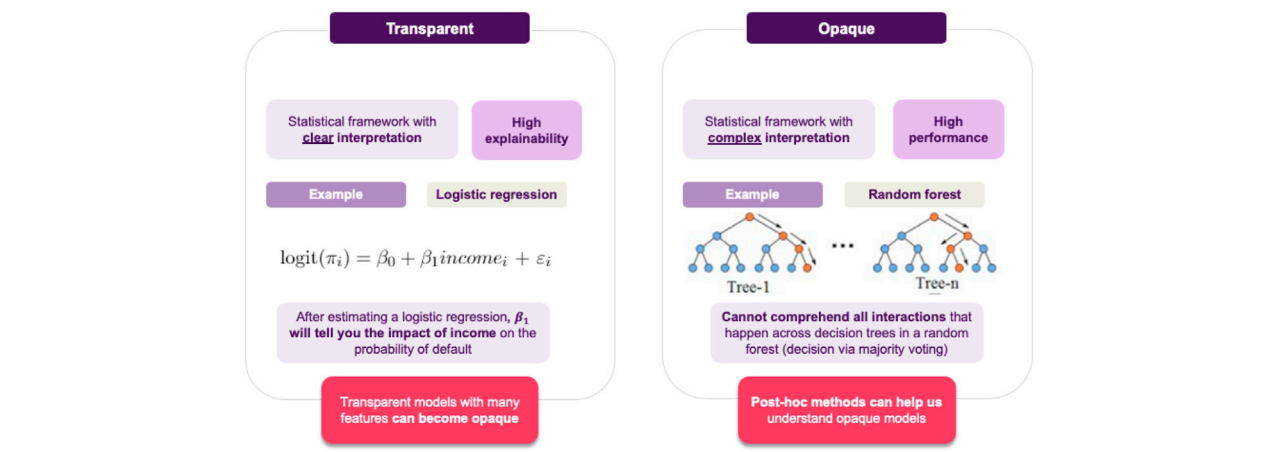


**FIGURE 4 |** Jane's choices: should she go for a transparent model or an opaque one?

Following these findings, the stakeholders are happy with both the model's performance and the degree of explainability. However, upon further inspection, they find out that there are some data points in the training dataset that are too noisy, probably not corresponding to actual data, but rather to instances that were included in the dateset by accident. They turn to Jane, in order to get some insights about how deleting these data points from the training dataset would affect the models behavior. Fortunately, deletion diagnostics show that omitting these instances would not affect the models performance, while they were able to identify some points that could significantly alter the decision boundary, too (**Figure 10**). All of these helped the stakeholder understand which training data points were more influential for the model.

Finally, as an extra layer of protection, the stakeholders ask Jane if it is possible to have a set of rules describing the model's behavior on a global scale, so they can inspect it to find out whether the model has picked up any undesired functioning. At this point, Jane thinks that they should utilize the Random Forest's structure, which is an ensemble of Decision Trees. This means, that they already consist of a large number of rules, so it

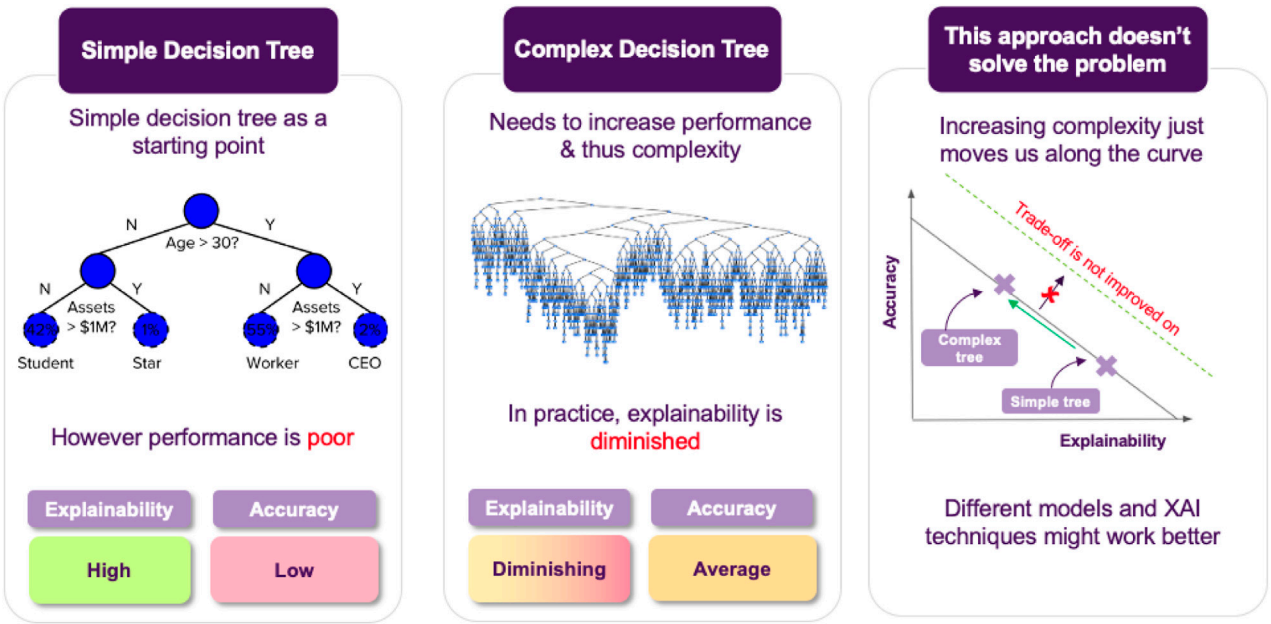


**FIGURE 5 |** As transparent models become increasingly complex they may lose their explainability features. The primary goal is to maintain a balance between explainability and accuracy. In cases where this is not possible, opaque models paired with post hoc XAI approaches provide an alternative solution.

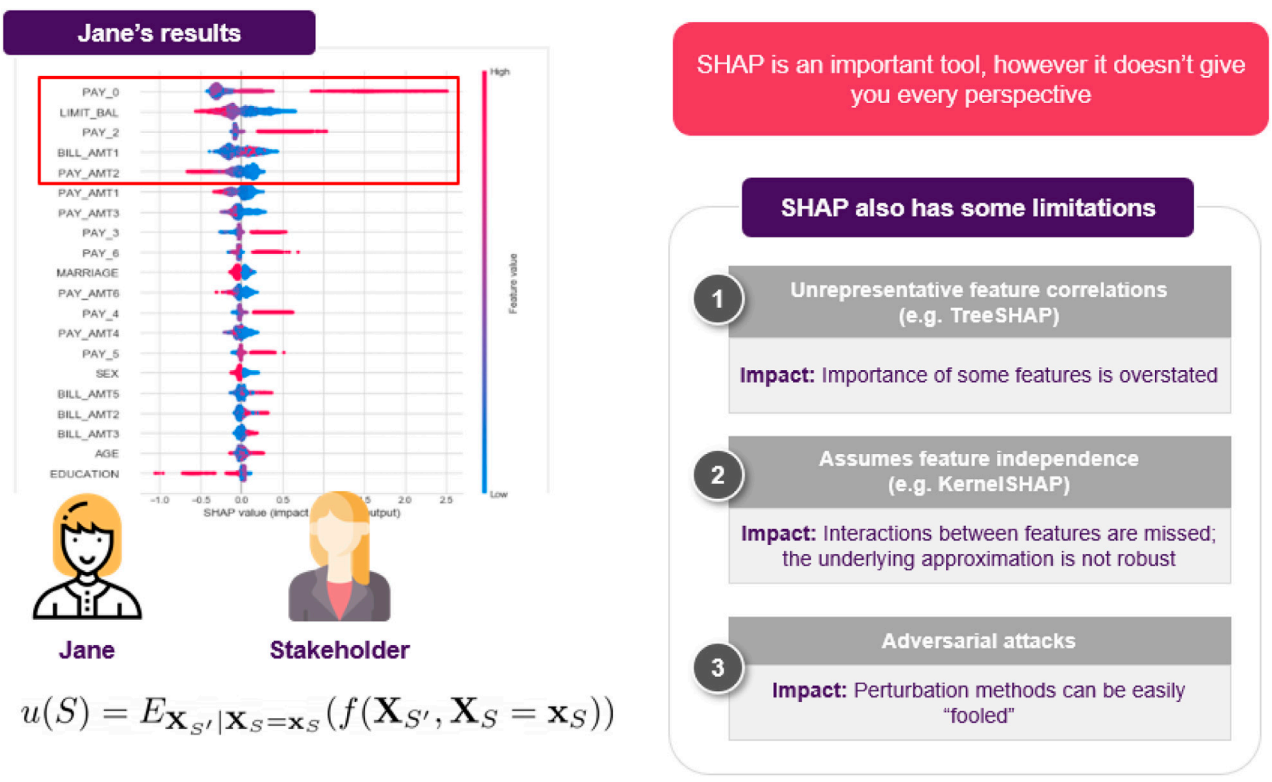


**FIGURE 6 |** Jane decides to use SHAP, but cannot resolve all of the stakeholder's questions. Its also worth noting that although SHAP is an important method for explaining opaque models, users should be aware of its limitations, often arising from either the optimization objective or the underlying approximation.

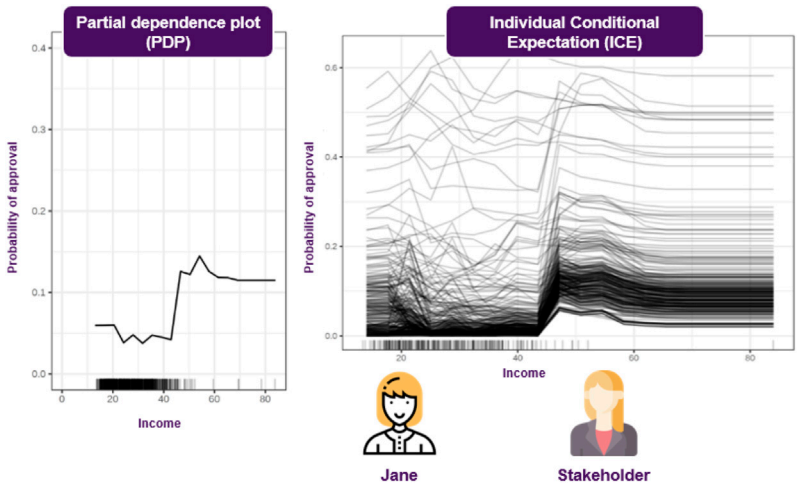


**FIGURE 7 |** Visualizations can facilitate understanding the model's reasoning, both on an instance and a global level. Most of these approaches make a set of assumptions, so choosing the appropriate one depends on the application.

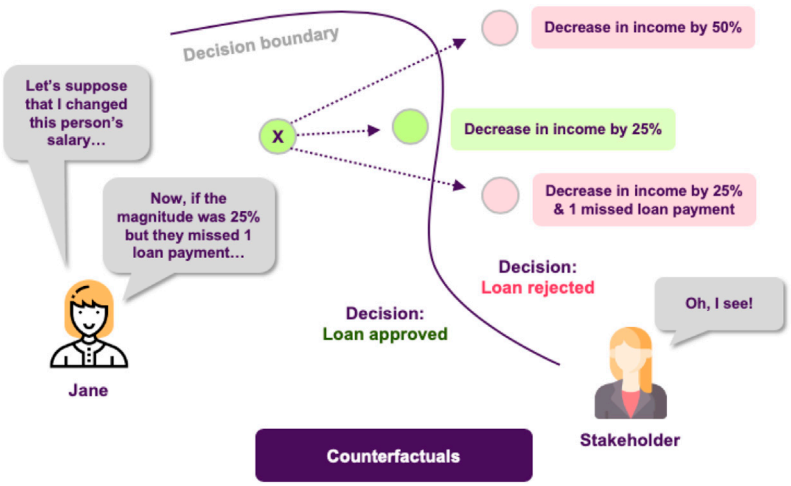


**FIGURE 8 |** Counterfactuals produce a hypothetical instance, representing a minimal set of changes of the original one, so the model classifies it in a different category.

makes sense to go for an approach that is able to extract the more robust ones, such as in Trees (**Figure 11**).

The above example showcases how different XAI approaches can be applied to a model to answer various types of questions. Furthermore, the last point highlights an interesting distinction, as SHAP, anchors and counterfactuals that are model agnostic, while in Trees are model-specific, utilizing the model's architecture to produce explanations. There are some points to note here (cf. **Figure 12**): model agnostic techniques apply to any model, and so if benchmarking a whole range of models, inspecting their features, model agnostic methods offer consistency in interpretation. On the other hand, since these approaches have to be very flexible, a significant amount of assumptions and approximations may be made, possibly resulting in poor estimates or undesired side-effects, such as susceptibility to adversarial attacks (Slack et al., 2020). Model-specific could also facilitate developing more efficient algorithms or custom flavored explanations, based on the model's characteristics.

Another factor to take into consideration has to do with the libraries, since model-agnostic approaches are usually widely used and compatible with various popular libraries, whereas model-specific ones are emerging and fewer, with possibly only academic libraries being available. Overall, attempting to use a larger set of XAI methods allows for deeper inquiry (cf. **Figure 13**).

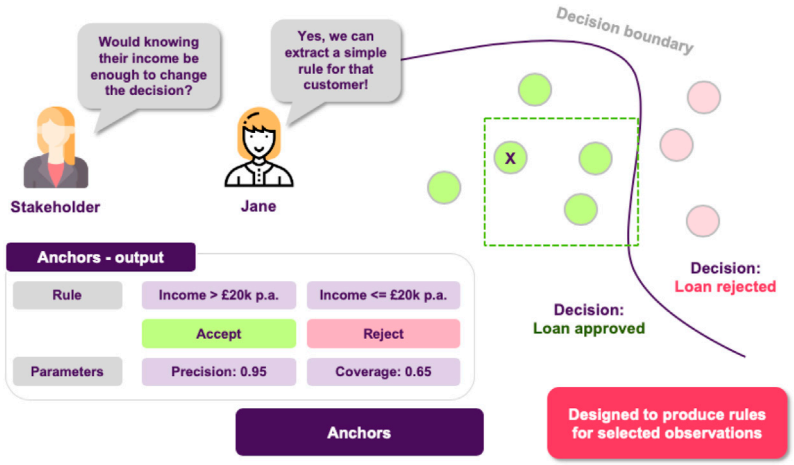


**FIGURE 9 |** Local explanations as rules. High precision means that the rule is robust and that similar instances will get the same outcome. High coverage means that large number of the points satisfy the rule's premises, so the rule "generalizes" better.

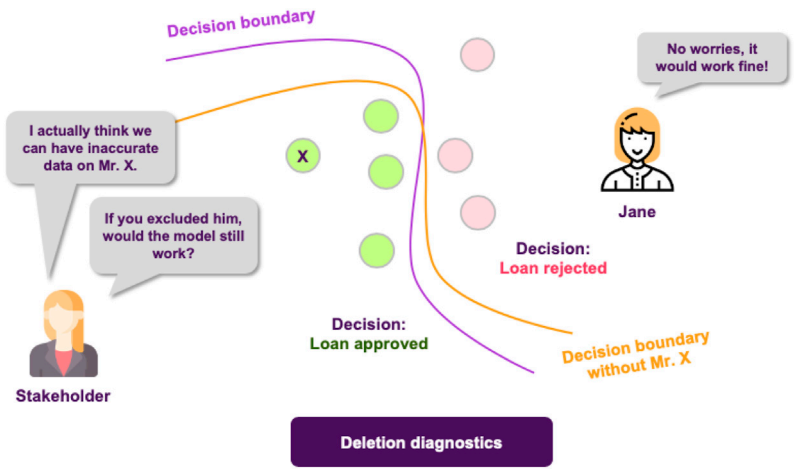


**FIGURE 10 |** The quality of a ML model is vastly affected by the quality of the data it is trained on. Finding influential points that can, for example, alter the decision boundary or encourage the model to take a certain decision, contributes in having a more complete picture of the model's reasoning.

These insights are summarized in terms of a "cheat sheet." **Figure 14** discuss a sample pipeline in terms of approaching explainability for machine learning, and **Figure 15**[7] and **Figure 16**[8] discusses possible methods.

## 12 FUTURE DIRECTIONS

This survey offers an introduction in the various developments and aspects of explainable machine learning. Having said that, XAI is a relatively new and still developing field, meaning that there are many open challenges that need to be considered, not all of them lying on the technical side. Of course, generating accurate and meaningful explanations is important, but communicating them in an effective manner to a diverse audience, is equally important. In fact, a recent line of work addressing the interconnection between explanations and communication has already emerged within the financial sector.

Considering the risks of financial investments, it should probably come as no surprise that the importance of employing XAI when using opaque ML models in finance, has

[7]Links to packages (in Python and R): shap.readthedocs.io/en/latest/, cran.r-project.org/web/packages/shapper/index.html, scikit-learn.org/stable/modules/partial_dependence.html, bgreenwell.github.io/pdp/articles/pdp.html, docs.seldon.io/projects/alibi/en/latest/

[8]Links to packages (in Python and R): docs.seldon.io/projects/alibi/en/latest/, github.com/viadee/anchorsOnR, www.statsmodels.org/stable/generated/statsmodels.stats.outliers_influence.OLSInfluence.html, www.rdocumentation.org/packages/stats/versions/3.6.2, github.com/IBCNServices/GENESIM/blob/master/constructors/inTrees.py, cran.r-project.org/web/packages/inTrees/index.html

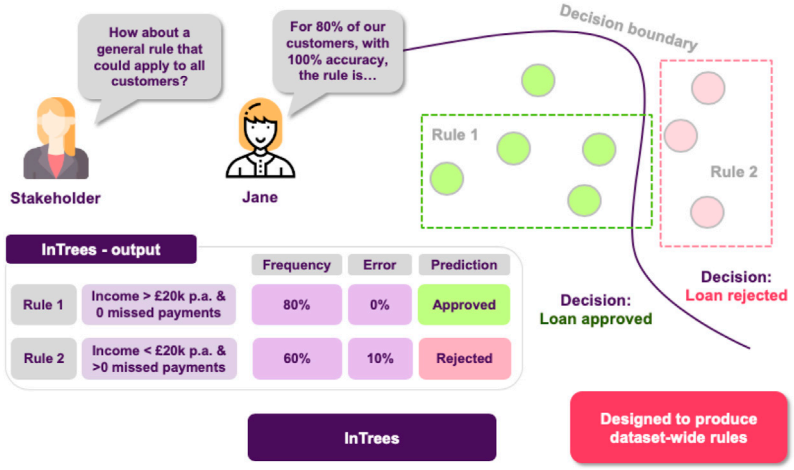


**FIGURE 11 |** Extracting rules from a random forest. Frequency of a rule is defined as the proportion of data instances satisfying the rule condition. The frequency measures the popularity of the rule. Error of a rule is defined as the number of incorrectly classified instances determined by the rule. So she is able to say that for 80% of the customers with 100% accuracy (ie. 0% error), when income >20 k and there are zero missed payments, the application is approved.

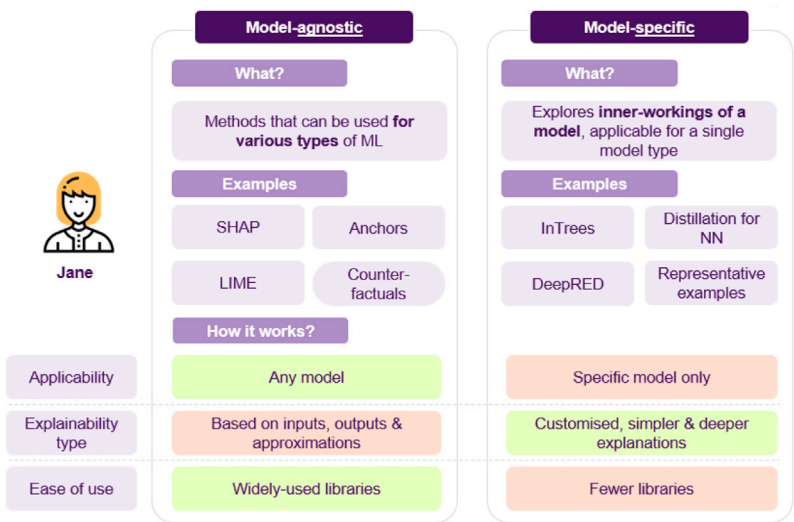


**FIGURE 12 |** A short comparison of model agnostic vs. model specific approaches.

been already identified (FSB, 2017; Croxson et al., 2019; Joseph, 2019). However, the traits that render an explanation satisfying are not independent of the audience's characteristics and expectations. To this end, a series of recent papers address this exact question (van den Berg and Kuiper, 2020; Langer et al., 2021), highlighting the need to consider the point of view of the various stakeholders. As a consequence, explanations should be tailored to the specific audience they are intended for, aiming at conveying the necessary information in a clear way. This observation has naturally led to the development of XAI approaches that specifically target financial applications (Philippe et al., 2019; Bussmann et al., 2020; Misheva et al., 2021), but further research could lead to more significant advances.

It is interesting to note that approaches from the broader AI community, eg (Kulkarni et al., 2019; Chakraborti et al., 2019), mentioned in **Section 3**, also attempt to tackle this problem, but by means of a formal approach. Indeed, in the area of human-aware AI, there is an increasing focus on explicitly modeling the mental state, beliefs and expectations of the user and reconciling that with the system's model of the world. See, for example, discussion in (Kambhampati, 2020). Admittedly, such frameworks do not yet consider general stakeholder concerns in complex environments and so it would be interesting to see if such frameworks might eventually be helpful in areas such as finance.

Another non-technical matter that has is getting increasing attention is concerned with the incorporation of XAI in

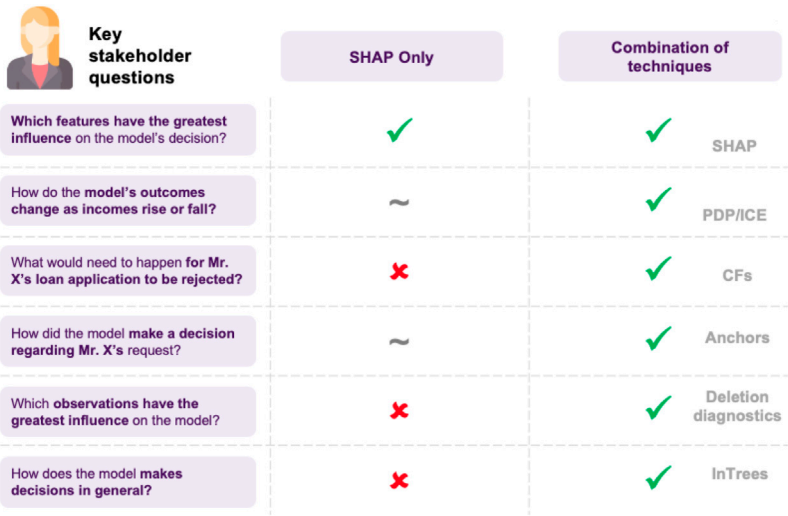


**FIGURE 13 |** A list of possible questions of interest when explaining a model. This highlights the need for combining multiple techniques together and that there is no catch-all approach.

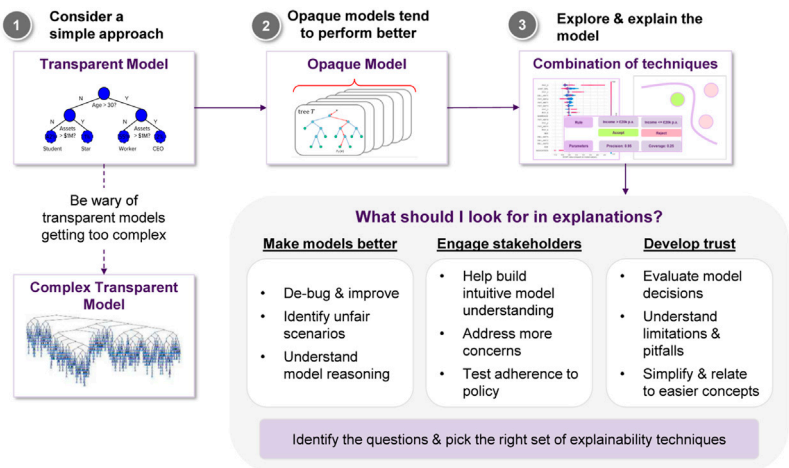


**FIGURE 14 |** A sample pipeline, that is, a "cheat sheet" of sorts for approaching explainability.

regulatory frameworks. The European GDPR [EU Regulation (EU), 2016] regulation can be seen as an early attempt toward this objective, since it requires automated models to be able to provide meaningful information about their rationale, which in turn motivated additional XAI related research. Apart from GDPR, the European Commission has recently published an ethics guideline for trustworthy AI (High-Level Expert Group on AI, 2019), where the need of being able to explain a model's decisions is deemed an essential requirement for establishing trust between human users and AI systems. On top of that, an even more recent paper by the European Commission (European Commission, 2020) has touched upon this issue, and how explainability can be a key element for developing a future regulatory framework. However, the regulatory integration of XAI is still an ongoing process, where additional interdisciplinary research is needed in order to develop a framework that fulfills all the necessary requirements.

When it comes to the technical side of XAI, there are many research and operational open problems that need to be considered, as research progresses (cf. **Figure 17**).

One of the first things that comes to mind is related to the way that different explanation types fit with each other. If we take a close look at the presented approaches, we will find out that while there is some overlap between the various explanation types, for the most part they appear to be segmented, each one addressing a different question. Moreover, there seems to be no clear way of combining them to produce a more complete explanation. This hinders the development of pipelines that aim at automating explanations, or even reaching an agreement on how a complete explanation should look like.

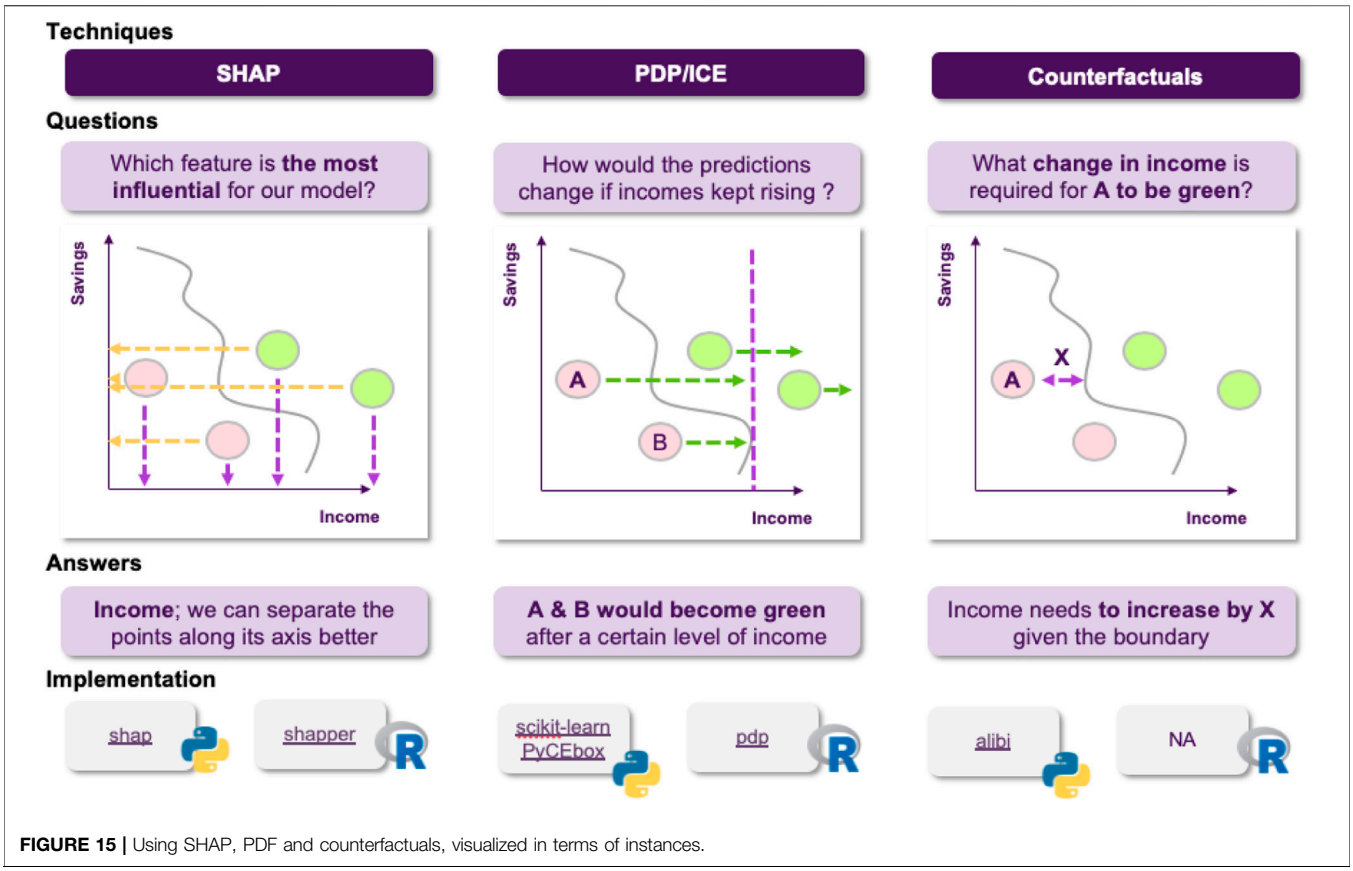


**FIGURE 15 |** Using SHAP, PDF and counterfactuals, visualized in terms of instances.

On a more practical level, there are only a few XAI approaches that come with efficient implementations. This could be justified by the fact that the field is still young and emerging, but it impedes the deployment of XAI in large scale applications, nonetheless.

Another aspect that could receive more attention in the future, is developing stronger model-specific approaches. The advantage of exploring this direction is that the resulting approaches would be able to utilize a model's distinct features to produce explanations, probably improving fidelity, as well as allowing to better analyze the model's inner workings, instead of just explaining its outcome. Furthermore, a side note related to the previous point is that this would probably facilitate coming up with efficient algorithmic implementations, since the new algorithms would not rely on costly approximations.

This last point leads to a broader issue that needs to be resolved, which is building trust toward the explanations themselves. As we mentioned before, recent research has showcased how a number of popular, widely used, XAI approaches are vulnerable to adversarial attacks (Slack et al., 2020). Information like that raises questions about whether the outcome of a XAI technique should be trusted or it has been manipulated. Furthermore, other related issues about the fitness of some of the proposed techniques can be found in the literature (Kumar I. E. et al., 2020). A promising way to address robustness issues is through exploring additional ways of establishing connections between XAI and statistics, opening up the door for utilizing a wide array of well studied tools.

Another line of research that has recently gained traction is about designing hybrid models, combining the expressiveness of opaque models with the clear semantics of transparent models, as in (Munkhdalai et al., 2020), where linear regression is combined with neural networks, for example. This direction could not only help bridge the gap between opaque and transparent models, but could also aid the development of state-of-the-art performing explainable models.

Finally, as XAI matures, notions of causal analysis should be incorporated to new approaches (Pearl, 2018; Miller, 2019). This is already a major driver in fundamental problems in other areas, such as fairness and bias in machine learning (Dwork et al., 2012; Kusner et al., 2017), so we should expect it to play an integral part in the future of the XAI literature.

## AUTHOR CONTRIBUTIONS

VB supervised the writing of the manuscript. Both authors contributed to the final version of the manuscript.

## ACKNOWLEDGMENTS

VB was partly supported by a Royal Society University Research Fellowship. IP was partly supported by the EPSRC grant Towards Explainable and Robust Statistical AI: A Symbolic

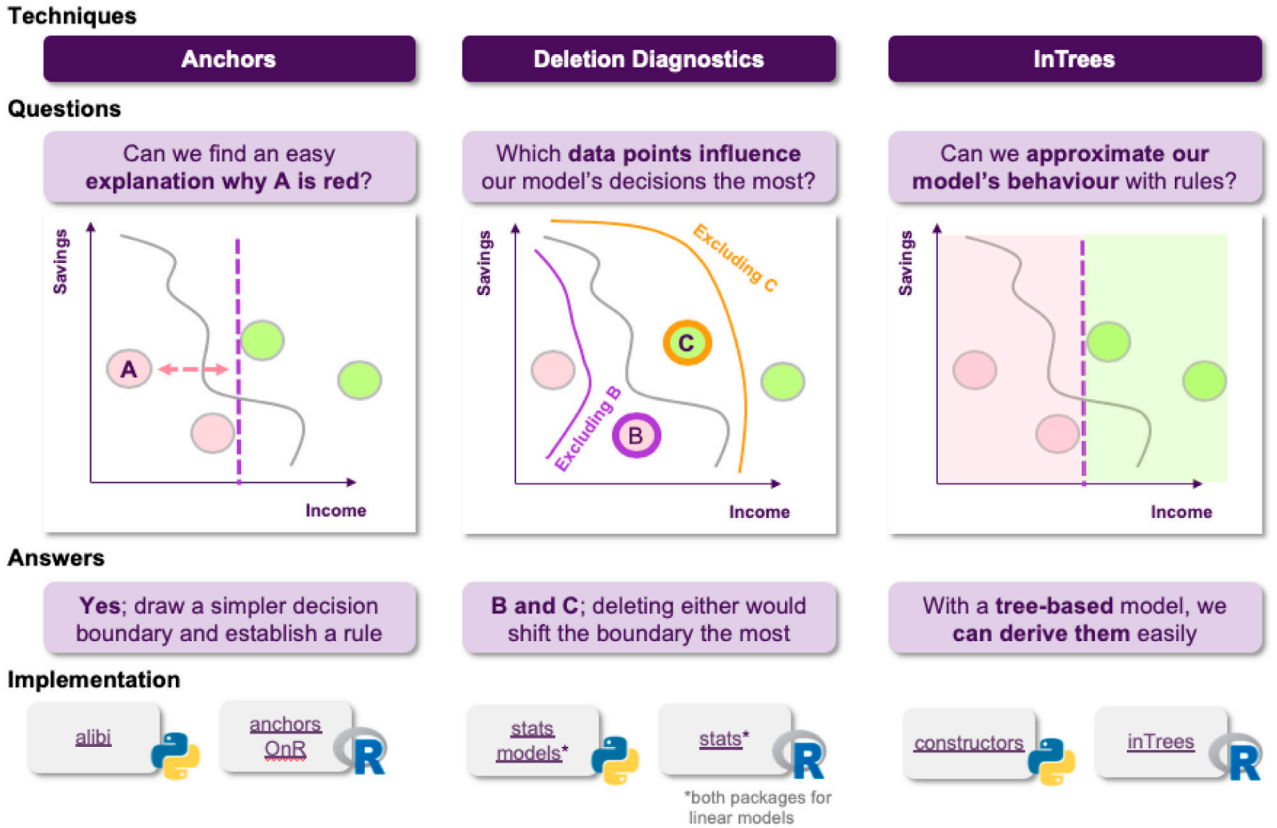


**FIGURE 16 |** Using anchors, deletion diagnostics and intrees, visualized in terms of instances.

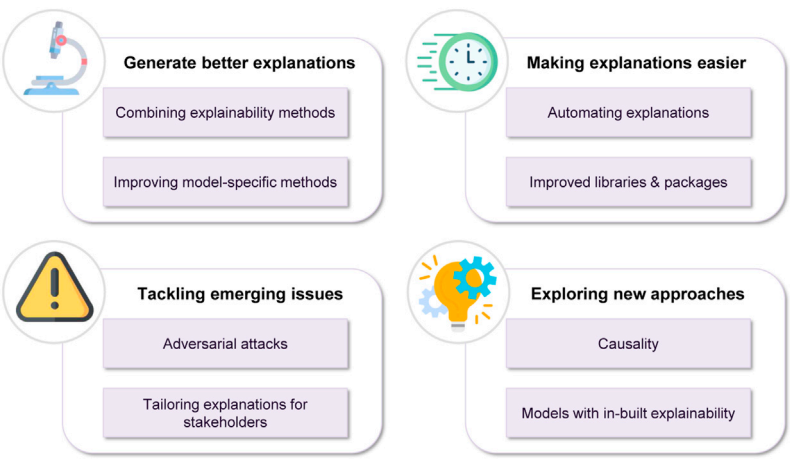


**FIGURE 17 |** Possible avenues for XAI research.

Approach. The authors acknowledge the financial support received by NatWest Group. This work was carried out in collaboration with University of Edinburgh's Bayes Centre and NatWest Group. We are especially grateful to Peter Gostev from the Data Strategy and Innovation team as well as a wide range of teams throughout Data and Analytics function at NatWest Group who provided insights on industry use cases, key issues faced by financial institutions as well as on the applicability of machine learning techniques in practice.

## REFERENCES


Adebayo, J., and Kagal, L. (2016). *Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models*, FATML Workshop 2016, New York, NY.

Agrahari, R., Foroushani, A., Docking, T. R., Chang, L., Duns, G., Hudoba, M., Karsan, A., and Zare, H. (2018). Applications of Bayesian Network Models in Predicting Types of Hematological Malignancies. Scientific Reports.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2019). Explainable Artificial Intelligence (Xai): Concepts, Taxonomies, Opportunities and Challenges toward Responsible Ai. *arXiv preprint arXiv:1910.10045*.

Augasta, M. G., and Kathirvalavakumar, T. (2012). Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems. Heidelberg, Germany. Neural Processing Letters.

Auret, L., and Aldrich, C. (2012). Interpretation of Nonlinear Relationships between Process Variables by Use of Random Forests. *Minerals Eng.* 35, 27–42. doi:10.1016/j.mineng.2012.05.008

Bastani, O., Kim, C., and Bastani, H. (2017). *Interpretability via Model Extraction.ArXiv, abs/1706.09773*

Baum, L. E., and Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Ann. Math. Statist.* 37 (6), 1554–1563. doi:10.1214/aoms/1177699147

Belle, V. (2019). Abstracting Probabilistic Models: A Logical Perspective. Ninth International Workshop on Statistical Relational AI, StarAI 2020, New York, NY, February 2020. Through 07-02-2020.

Ben-Hur, A., Horn, D., Siegelmann, H., and Vapnik, V. N. (2001). *Support Vector Clustering*. Brookline, MA: Microtome Publishing.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. Proceedings of the fifth annual workshop on Computational learning theory – COLT '92, Pittsburgh, PA, July 1992.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks.

Bucila, C., Caruana, R., and Niculescu-Mizil, A. (2006). *Model Compression*. New York, NY: Association for Computing Machinery.

Chakraborti, T., Sreedharan, S., Grover, S., and Kambhampati, S. (2019). Plan Explanations as Model Reconciliation. 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Daegu, Korea, March 2019. IEEE, 258–266.

Chastaing, G., Gamboa, F., and Prieur, C. (2012). *Generalized Hoeffding-Sobol Decomposition for Dependent Variables - Application to Sensitivity Analysis*. Beachwood, OH: Institute of Mathematical Statistics and the Bernoulli Society.

Che, Z., Purushotham, S., Khemani, R., and Liu, Y. (2016). Interpretable Deep Models for Icu Outcome Prediction. *AMIA Annu. Symp. Proc.* 2016, 371–380.

Chen, K., Hwu, T., Kashyap, H. J., Krichmar, J. L., Stewart, K., Xing, J., et al. (2020). Neurorobots as a Means toward Neuroethology and Explainable AI. *Front. Neurorobot.* 14, 570308. doi:10.3389/fnbot.2020.570308

Chicco, D., Sadowski, P., and Baldi, P. (2014). Deep Autoencoder Neural Networks for Gene Ontology Annotation Predictions. Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '14, New York, NY, September 2014. New York, NY: Association for Computing Machinery, 533–540.

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics* 19 (1), 15–18. doi:10.2307/1268249

Cortes, C., and Vapnik, V. (1995). *Support-vector Networks*. Cham, Switzerland: Springer Nature Switzerland AG.

Cortez, P., and Embrechts, M. J. (2011). Opening Black Box Data Mining Models Using Sensitivity Analysis. *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, New York, NY: Institute of Electrical and Electronics Engineers, 341–348.

Cortez, P., and Embrechts, M. J. (2013). Using Sensitivity Analysis and Visualization Techniques to Open Black Box Data Mining Models. *Inf. Sci.* 225, 1–17. doi:10.1016/j.ins.2012.10.039

Craven, M., and Shavlik, J. (1999). *Rule Extraction: Where Do We Go from Here*, 99. Madison, Wisconsin: University of Wisconsin Machine Learning Research Group working Paper.

Craven, M. W., and Shavlik, J. W. (1994). "Using Sampling and Queries to Extract Rules from Trained Neural Networks," in Machine Learning Proceedings. Editors W. W. Cohen and H. Hirsh (San Francisco (CA): Morgan Kaufmann), 37–45. doi:10.1016/b978-1-55860-335-6.50013-1

Croxson, K., Bracke, P., and Jung, C. (2019). *Explaining Why the Computer Says 'no'*. London, UK: FCA-Insight.

Dasgupta, D. (1999). *Artificial Immune Systems and Their Applications*. Berlin, Germany: Springer-Verlag Berlin Heidelberg.

Datta, A., Sen, S., and Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. *IEEE Symposium on Security and Privacy (SP)*. New York, NY: Institute of Electrical and Electronics Engineers, 598–617.

Deng, H. (2014). Interpreting Tree Ensembles with Intrees. *arXiv:1408.5456*.

Doshi-Velez, F., and Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*.

Drucker, H., Burges, C. C., Kaufman, L., Smola, A. J., and Vapnik, V. N. (1996). Support Vector Regression Machines. *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through Awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS '12. 214–226. New York, NY, USA: Association for Computing Machinery.

EU, Regulation (EU) (2016). 2016/679–general Data protection Regulation (GDPR). Brussels, Belgium: European Union.

European Commission (2020). *On Artificial Intelligence – A European Approach to Excellence and Trust*. Brussels, Belgium: European Union.

Fawcett, Tom. (2006). *An Introduction to ROC Analysis*. Pattern Recognition Letters. New York, NY, Elsevier Science Inc.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Statist.* 29 (5), 1189–1232. doi:10.1214/aos/1013203451

Friedman, J. H., and Meulman, J. J. (2003). Multiple Additive Regression Trees with Application in Epidemiology. *Statist. Med.* 22 (9), 1365–1381. doi:10.1002/sim.1501

FSB (2017). *Artificial Intelligence and Machine Learning in Financial Services–Market Developments and Financial Stability Implication, Technical Report* (Basel, Switzerland: . Financial Stability Board).

Fu, Li. Min. (1994). Rule Generation from Neural Networks. *IEEE Trans. Syst. Man. Cybern.* 24 (8), 1114–1124. doi:10.1109/21.299696

Geiger, D., Verma, T., and Pearl, J. (1990). *Identifying independence in Bayesian Networks*. Hoboken, NJ: Networks.

Giudici, Paolo., and Raffinetti, Emanuela. Shapley-Lorenz eXplainable Artificial Intelligence. *Expert Systems with Applications*. Amsterdam, Netherlands: Elsevier.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2013). *Peeking inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation*. Oxfordshire, United Kingdom: Taylor & Francis.

Gunning, D. (2017). Explainable Artificial Intelligence (Xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence 2.

Hara, S., and Hayashi, K. (2016). *Making Tree Ensembles Interpretable*. Lanzarote, Spain: Proceedings of Machine Learning Research.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning*. New York, NY: Springer, 587–604. doi:10.1007/978-0-387-84858-7_15

Henelius, A., Puolamaki, K., Bostrom, H., Asker, L., and Papapetrou, P. (2014). A Peek into the Black Box: Exploring Classifiers by Randomization. *Data Mining Knowledge Discov.* 28 (5-6), 1503–1529. doi:10.1007/s10618-014-0368-8

Henelius, A., Puolamäki, K., and Ukkonen, A. (2017). *Interpreting Classifiers through Attribute Interactions in Datasets*. Norwell, MA: Kluwer Academic Publishers.

High-Level Expert Group on AI (2019). *Ethics Guidelines for Trustworthy AI*. Brussels, Belgium: European Commission.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the Knowledge in a Neural Network. *NIPS Deep Learning and Representation Learning Workshop*.

Hruschka, E. R., and Ebecken, N. F. (2006). Extracting Rules from Multilayer Perceptrons in Classification Problems: A Clustering-Based Approach. *Neurocomputing* 70 (1), 384–397. Neural Networks. doi:10.1016/j.neucom.2005.12.127

Johansson, U., König, R., and Niklasson, L. (2004a). *The Truth Is in There - Rule Extraction from Opaque Models Using Genetic Programming*. Miami Beach, FL: AAAI Press.

Johansson, U., Niklasson, L., and König, R. (2004b). *Accuracy vs. Comprehensibility in Data Mining Models*. Mountain View, Calif: International Society of Information Fusion.

John, H. S. (2017). *Probabilistic Program Abstractions*.

Joseph, A. (2019). *Shapley Regressions: A Framework for Statistical Inference in Machine Learning Models, Staff Working Paper No. 784*. London, United Kingdom: Bank of England.

Kahramanli, H., and Allahverdi, N. (2009). Rule Extraction from Trained Adaptive Neural Networks Using Artificial Immune Systems. *Expert Syst. Appl.* 36 (2, Part 1), 1513–1522. doi:10.1016/j.eswa.2007.11.024

Kambhampati, S. (2020). *Challenges of Human-Aware AI Systems*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence.

Kenett, R. S. (2012). *Applications of Bayesian Networks*. Amsterdam, Netherlands: Elsevier.

Kim, B., Rudin, C., and Shah, J. (2014). The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, December 2014, 2. Cambridge, MA, USA: MIT Press, NIPS'141952–141960.

Kindermans, P.-J., Schütt, K. T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., et al. (2017). *Learning How to Explain Neural Networks: Patternnet and Patternattribution*. Vancouver, Canada: International Society of the Learning Sciences.

Koh, P. W., and Liang, P. (2017). Understanding Black-Box Predictions via Influence Functions. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, August 2017, 70. (JMLR.org), 1885–171894.ICML'17.

Konig, R., Johansson, U., Niklasson, L., and rex, G. (2008). A Versatile Framework for Evolutionary Data Mining. In IEEE International Conference on Data Mining Workshops, pages 971–974.

Koshevoy, G., and Mosler, K. (1996). The Lorenz Zonoid of a Multivariate Distribution. *J. Am. Stat. Assoc.* 91, 873–882. doi:10.1080/01621459.1996.10476955

Krishnan, S., and Wu, E. (2017). Palm: Machine Learning Explanations for Iterative Debugging. Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics, Chicago, IL, May 2017. (New York, NY, USA: HILDA'17Association for Computing Machinery).

Kulkarni, A., Zha, Y., Chakraborti, T., Vadlamudi, S. G., Zhang, Y., and Kambhampati, S. (2019). Explicable Planning as Minimizing Distance from Expected BehaviorProceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, Montreal, QC, May 2019. (International Foundation for Autonomous Agents and Multiagent Systems), 2075–2077.

Kumar, I. E., Scheidegger, C., Venkatasubramanian, S., and Friedler, S. (2020b). Shapley Residuals: Quantifying the Limits of the Shapley Value for Explanations. *ICML Workshop on Workshop on Human Interpretability in Machine Learning*, July 2020. (WHI).

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. (2020a). *Problems with Shapley-Value-Based Explanations as Feature Importance Measures*, Proceedings of Machine Learning Research, July 2020.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). "Counterfactual Fairness," in *Advances in Neural Information Processing Systems*. Editors I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Long Beach, CA: Curran Associates, Inc.), 30, 4066–4076.

Kyrimi, E., Mossadegh, S., Tai, N., and Marsh, W. (2020). An Incremental Explanation of Inference in Bayesian Networks for Increasing Model Trustworthiness and Supporting Clinical Decision Making. *Artificial Intelligence in Medicine*. Amsterdam, Netherlands: Elsevier.

Langer, Markus., Oster, Daniel., Speith, Timo., Hermanns, Holger., Kastner, Lena., Schmidt, Eva., et al. (2021). *What Do We Want from Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research*. Amsterdam, Netherlands: Artificial Intelligence.

Lundberg, S. M., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, December 2017, NIPS'17. NY, USA: Red HookCurran Associates Inc, 4768–4777.

Mashayekhi, M., and Gras, R. (2015). "Rule Extraction from Random forest: the Rf+hc Methods," in *Advances in Artificial Intelligence*. Editors D. Barbosa and E. Milios (Cham: Springer International Publishing), 223–237. doi:10.1007/978-3-319-18356-5_20

Merrick, L., and Taly, A. (2019). *The Explanation Game: Explaining Machine Learning Models with Cooperative Game Theory*. Dublin, Ireland: Machine Learning and Knowledge Extraction.

Micaelli, P., and Storkey, A. J. (2019). "Zero-shot Knowledge Transfer via Adversarial Belief Matching," in *Advances in Neural Information Processing Systems*. Editors H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett (Vancouver, BC: Curran Associates, Inc.), 32, 9547–9557.

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artif. Intelligence* 267, 1–38. doi:10.1016/j.artint.2018.07.007

Misheva, Branka. Hadji., Osterrieder, Joerg., Ali, Hirsa., Kulkarni, Onkar., and Lin, Stephen. Fung. (2021). Explainable AI in Credit Risk Management. *arxiv Quantitative Finance*.

Molnar, C. (2020). *Interpretable Machine Learning*. Morrisville, NC: Lulu. com.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining Nonlinear Classification Decisions with Deep taylor Decomposition. *Pattern Recognition* 65, 211–222. doi:10.1016/j.patcog.2016.11.008

Munkhdalai, L., Munkhdalai, T., and Ryu, K. H. (2020). A Locally Adaptive Interpretable Regression. *arXiv*. [Epub ahead of print].

Owen, A. B., and Prieur, C. (2017). On Shapley Value for Measuring Importance of Dependent Inputs. *SIAM/ASA J. Uncertainty Quantification* 5, 986–1002. doi:10.1137/16m1097717

Owen, Art. B. (2013). *Variance Components and Generalized Sobol' Indices*. Philadelphia, PA: SIAM/ASA Journal on Uncertainty Quantification.

Özbakundefinedr, L., Baykasoundefinedlu, A., and Kulluk, S. (2010). A Soft Computing-Based Approach for Integrated Training and Rule Extraction from Artificial Neural Networks: Difaconn-Miner. *Appl. Soft Comput.* 10 (1), 304–317. doi:10.1016/j.asoc.2009.08.008

Palczewska, A., Palczewski, J., Robinson, R. M., and Neagu, D. (2013). Interpreting Random forest Classification Models Using a Feature Contribution Method. *ArXiv, abs/1312*.1121.

Pearl, J. Theoretical Impediments to Machine Learning with Seven sparks from the Causal Revolution. *arXiv preprint arXiv:1801.04016*, (2018).

Petkovic, D., Altman, R., Wong, M., and Vigil, A. (2018). Improving the Explainability of Random forest Classifier - User Centered Approach. *Pacific Symposium on Biocomputing*. Kohala Coast, Hawaii: World Scientific Publishing Company.

Philippe, Bracke., Datta, Anupam., Jung, Carsten., and Sen, Shayak. (2019). *Machine Learning Explainability in Finance: An Application to Default Risk Analysis*. London, United Kingdom: Bank of England.

Ribeiro, M. T., Singh, S., and Anchors, C. Guestrin. (2018). *High-precision Model-Agnostic Explanations*. New Orleans Riverside, New Orleans: AAAI Press.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, San Francisco, CA, August 2016. (New York, NY, USA: Association for Computing Machinery), 1135–1144.

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach Intell.* 1 (5), 206–215. doi:10.1038/s42256-019-0048-x

Saad, E. W., and Wunsch, D. C. (2007). Neural Network Explanation Using Inversion. *Neural Networks* 20 (1), 78–93. doi:10.1016/j.neunet.2006.07.005

Sato, M., and Tsukimoto, H. (2001). Rule Extraction from Neural Networks via Decision Tree Induction. *IJCNN'01. International Joint Conference On Neural Networks*. Proceedings (Cat. No.01CH37222), Washington, DC, July 2001. 3, 1870–1875.

Shapley, L. S. (1952). *A VALUE FOR N-PERSON GAMES*. Defense Technical Information Center. Santa Monica, CA: RAND Corporatio.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). "Learning Important Features through Propagating Activation Differences,"Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, Sydney, NSW, August 2017. Editors D. Precup and Y. W. Teh (Sydney, Australia: International Convention Centre PMLR), 3145–3153.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling Lime and Shap: Adversarial Attacks on Post Hoc Explanation Methods. AAAI/ACM

Conference on Artificial Intelligence, New York, NY, February 2020. Ethics, and Society (AIES).

Song, Eunhye., and Barry, L. (2016). *Nelson, and Jeremy Staum, Shapley Effects for Global Sensitivity Analysis: Theory and Computation*. Philadelphia, PA: SIAM/ASA Journal on Uncertainty Quantification.

Strumbelj, E., and Kononenko, I. (2010). An Efficient Explanation of Individual Classifications Using Game Theory. *J. Mach. Learn. Res.* 11, 1–18. doi:10.1145/1756006.1756007

Su, G., Wei, D., Kush, R., and Malioutov, D. (2016). *Interpretable Two-Level Boolean Rule Learning for Classification*. 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016). New York, NY, United States.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning-, Sydney, NSW, August 2017. Volume 70, pages 3319–3328. JMLR.org.ICML'17.

Tan, H. F., Hooker, G., and Wells, M. T. (2016). Tree Space Prototypes: Another Look at Making Tree Ensembles Interpretable. New Orleans, LA: *ArXiv, abs/1611.07115*,

Tan, S., Caruana, R., Hooker, G., and Lou, Y. (2017). *Distill-and-compare: Auditing Black-Box Models Using Transparent Model Distillation*. AIES 18.

Timmer, S. T., Meyer, J.-J. C., Prakken, H., Renooij, S., and Verheij, B. (2016). A Two-phase Method for Extracting Explanatory Arguments from Bayesian Networks. *International Journal of Approximate Reasoning*. New York, NY: Elsevier Science Inc.

Tolomei, G., Silvestri, F., Haines, A., and Lalmas, M. (2017). *Interpretable Predictions of Tree-Based Ensembles via Actionable Feature Tweaking*. New York, NY: Association for Computing Machinery.

Turner, R. (2016a). A Model Explanation System. IEEE 26th International Workshop on Machine Learning for Signal Processing, Salerno, Italy, September 2016. MLSP, 1–6.

Turner, R. (2016b). A Model Explanation System: Latest Updates and Extensions. *arXiv* [Epub ahead of print].

Van Assche, A., and Blockeel, H. (2007). "Seeing the forest through the Trees: Learning a Comprehensible Model from an Ensemble," in *Machine Learning: ECML 2007*. Editors J. N. Kok, J. Koronacki, R. L. d. Mantaras, S. Matwin, D. Mladenič, and A. Skowron (Berlin, Heidelberg: Springer Berlin Heidelberg), 418–429. doi:10.1007/978-3-540-74958-5_39

van den Berg, Martin., and Kuiper, Ouren., XAI in the Financial Sector. A Conceptual Framework for Explainable AI (XAI), Hogeschool Utrecht, Lectoraat Artificial Intelligence Version 1.1, (2020).

van den Oord, A., Dieleman, S., and Schrauwen, B. (2013). "Deep Content-Based Music Recommendation," in *Advances in Neural Information Processing Systems*. Editors C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc.), 26, 2643–2651.

Vapnik, V. N., and Lerner, A. Y. (1963). Pattern Recognition Using Generalized Portraits. Moscow, Russia: Automation and Remote Control.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the Gdpr. *Harv. J. L. Technol.* 31, 841–887. doi:10.2139/ssrn.3063289

Weld, D. S., and Bansal, G. (2019). The challenge of Crafting Intelligible Intelligence. *Commun. ACM* 62 (6), 70–79. doi:10.1145/3282486

Welling, S., Refsgaard, H., Brockhoff, P., and Clemmensen, L. (2016). Forest Floor Visualizations of Random Forests. *arXiv:1605.09196*.

Lipton, Z. C. The Mythos of Model Interpretability. *arXiv preprint arXiv:1606.03490*, (2016).

Zhou, Y., and Hooker, G. (2016). Interpreting Models via Single Tree Approximation. *Methodology: arXiv*.

Zilke, J. R., Loza Mencía, E., and Janssen, F. (2016). DeepRED - Rule Extraction from Deep Neural Networks. *Discovery Science*. Manhattan, NY: Springer International Publishing, 457–473. doi:10.1007/978-3-319-46307-0_29

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.